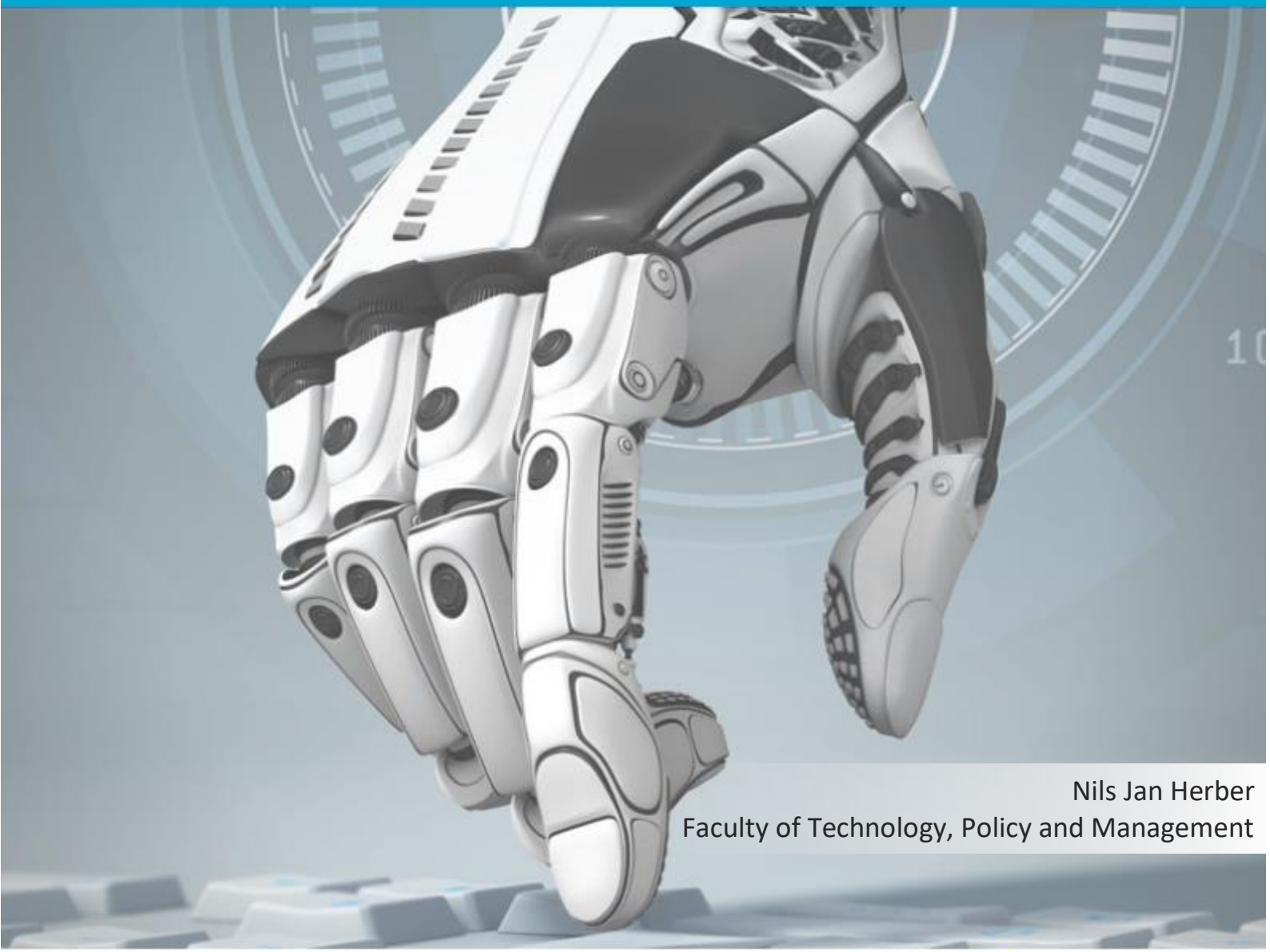
Towards Explaining Automated Credit Decisions

The design of an Explicability Assessment Framework (EAF) for Machine Learning Systems



Nils Jan Herber
Faculty of Technology, Policy and Management

TU Delft

EY Building a better working world

This page is intentionally left blank.

Towards Explaining Automated Credit Decisions

The design of an Explicability Assessment Framework (EAF)
for Machine Learning Systems

Master thesis submitted to Delft University of Technology

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in **Complex Systems Engineering and Management**

Faculty of Technology, Policy and Management

by

Nils Jan Herber

Student number: 4281489

To be defended in public on *24/10/2019*

Graduation committee

Chairperson	Prof. Dr. M.J. van den Hoven	TU Delft	Ethics/Philosophy of Technology section
First Supervisor	Dr. F. Santoni de Sio	TU Delft	Ethics/Philosophy of Technology section
Second Supervisor	Dr. S. Cunningham	TU Delft	Policy Analysis section
Advisor	PhD candidate S. Robbins, MSc.	TU Delft	Ethics/Philosophy of Technology section

External supervisor

Manager	J. Schijven, MSc.	EY	Financial Accounting Advisory Services department

TU Delft

EY Building a better working world

This page is intentionally left blank.

Acknowledgements

Dear reader,

In front of you lies the result of six months of hard work on my most challenging assignment so far. This thesis is the final chapter of 6 years studying in Delft and the final hurdle for my MSc. degree in Complex Systems Engineering and Management at the Technology, Policy and Management Faculty. I will always look back with a smile at the most valuable, exciting and fun time of my life here. I cannot explain how much all the smart and fun people I've met, the valuable knowledge that I've gained and the once in a lifetime experiences mean to me.

I am very happy that I've got the opportunity to contribute to one of the most exciting challenges that are out here nowadays by pursuing this thesis research. My thesis hopefully reflects the importance of these challenges. I'm intrinsically interested and driven to somehow further contribute to this area and I will try to keep up with all the progressions going on in the fast-evolving research field of AI ethics. Beforehand, I would like to take this opportunity to thank a number of important people.

First of all, I want to thank my supervisors. First my first supervisor, Filippo Santoni de Sio. Thank you for your day-to-day advice and guidance with this project. It was a pleasure to work with him, and without his devotion and counseling, this result wasn't possible. Also, I want to thank my second supervisor, chairman, and advisor for their valuable feedback and guidance; Jeroen van den Hoven, Scott Cunningham, and Scott Robbins. Despite their busy schedules, they have always been willing to help me, and I genuinely thank all four of them for that. I hope that after my graduation we will come across again.

Next, I want to thank the whole EY FAAS department for the opportunity to combine this thesis project with a thesis internship. This really helped with my motivation at some miserable rainy days, and foremost having fun all along with this project! I as well want to thank all the respondents for the valuable interviews that they've participated in.

I cannot forget to thank my friends for always being there for me and providing me with pleasant distractions at the moments that I needed it the most. A small extra thank you to Maurits, a true AI Ethics enthusiastic as well, for all the creative insights, mental and physical boosts that you've given me at our meetings in Leiden during my thesis period.

Last, but not least, I want to thank my family, and especially my father, my mother and my sister, Maren, for always supporting me all these years. This page is way too small to express how much all your encouragement, support, and trust means to me. Without you, this all wasn't possible.

Until prospective encounters! I cannot wait to see what the future holds for me

Nils Jan Herber
Delft University of Technology
October 24, 2019

Executive Summary

Machine learning is one of the leading technological innovation areas for the financial services industry. The self-learning algorithms have the potential to reduce costs, improve efficacy, find- and create new business ventures and improve risk management. One major application area is the use of machine learning systems to predict the probability of default with credit underwriting and hence more precise credit risk assessment. Despite these advantages, there are also issues that go with the adoption of this innovation.

The machine learning models with the highest prediction-accuracy are often the least explicable (i.e. explainable). This could cause problems with the non-discovery of unfair or discriminatory decision-making, workings of the system or human autonomy (being actively involved in the decision process). It is essential for decisions that have significant legal and social implications (such as credit decisions, or loan approvals/denials) that these can be explained. The main regulations that ask for machine learning models to be explicable (i.e. able to be explained) are the General Data Protection Regulation (GDPR) and the Consumer Credit Directive (CCD). Moreover, there exists an ethical need among academia and in society that requires automated decisions to be explicable.

Within the exploration of scientific literature, it becomes clear that it lacks research on how to move from a high-level principle like explicability, towards a prospective assessment of a machine learning use case on this principle. In addition, most machine learning ethics literature takes a mono-disciplinary approach, but explicability assessment requires a multi-disciplinary perspective. Lastly, the literature lacks an assessment framework that can guide decision-makers within machine learning use cases, that is aligned with a multi-organizational development lifecycle.

In order to solve the formerly mentioned problems, the main research question has been formulated as follows: *"How can decision-makers prospectively assess machine learning applications within credit underwriting from the point of view of explicability?"* The research objective that follows from this is: "*To design a pragmatic prospective assessment framework that can guide decision-makers, within machine learning applications in European credit underwriting cases from the point of view of explicability"*.

This thesis adopted the structured step-wise approach of the Design Science Research Methodology (DSRM) complemented with the VSD (Value Sensitive Design) approach. VSD is incorporated in the objectives formulation of the DSRM. The first step of DSRM is the problem identification, that has

taken place in the earlier paragraphs of this summary. The second step is the formulation of the design objectives which were derived from the complexities to solve. Expert interviews were conducted to validate the objectives and subsequent design requirements. The following design objectives were formulated:

The framework:

A. Provides guidance for the decision-makers on using the framework
B. Helps with 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

In the design and development step, the objectives were transformed to design requirements for the framework. Hereafter, means to commit to these design requirements were formulated and with these means, the artifact (Explicability Assessment Framework) was designed.

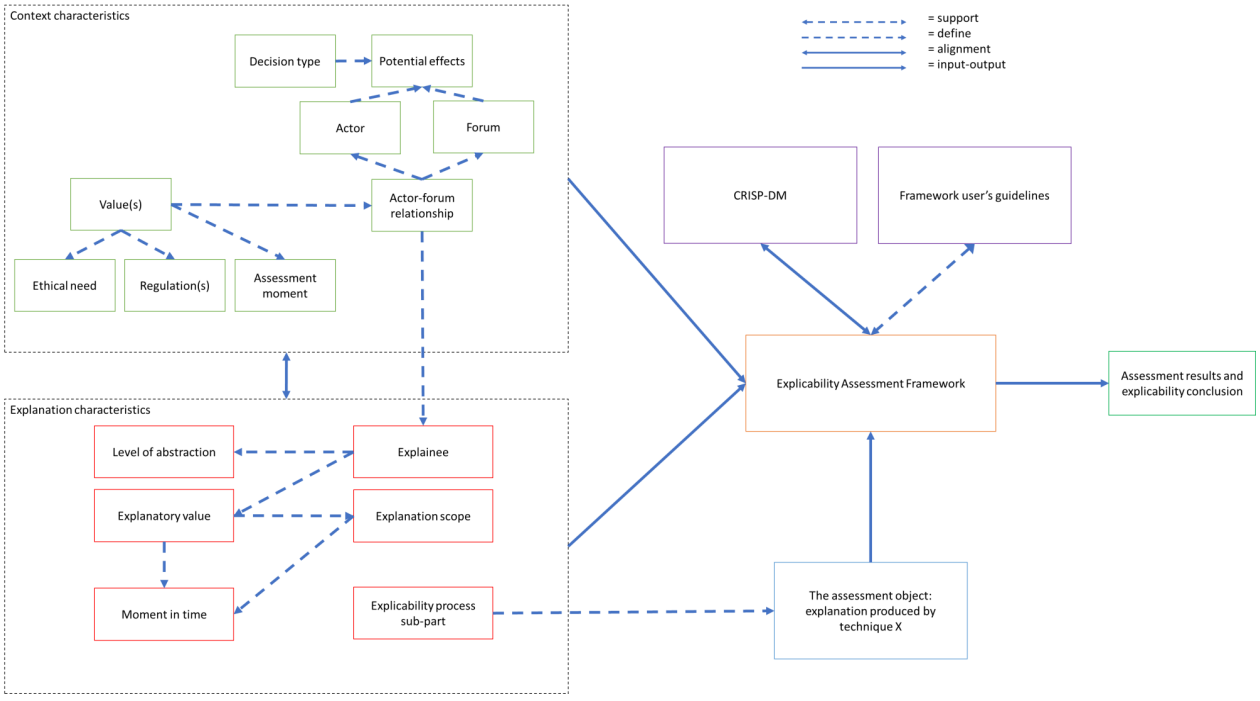


Figure 1: The Explicability Assessment Framework and the relationships (appendix 9.3.3.1. contains the framework enlarged)

Within the demonstration phase (step 4), the artifact is tested on use cases. We adopted two use cases from scientific papers that contain explanations to be assessed with the EAF: an *explanation by rule extraction* and a *counterfactual explanation*. Before being able to do this, the framework needs to be adapted to the context, with the context characteristics and explanation characteristics. This explanation scope has the following specifications within this thesis: *explicability process sub-part* is textual explanation product, *explainee* is layperson, *level of abstraction* is consumer-level, *explanatory value* is justification, *explanation scope* is local and *moment in time* is ex-post.

Step 5 concerns the evaluation of the artifact and the demonstration phase on the design objectives as defined before: to what extent did the artifact accomplish the formulated design objectives? It shows that the objectives are mainly achieved, with some having specific prerequisites and conditions.

These, in their turn, form the limitations of the framework. First, the framework covers a modest part of the full explicability spectrum to cover. Second, the framework does not include specific instructions on the choices for intelligibility types and evidence roles (in the case of justificatory value). Further, the soundness and conciseness factors of the framework do not have a clear threshold. Fourth, the usability of the framework for the proposed users' needs to be empirically validated. Last, the general knowledge level within the application area of laypersons has not been empirically investigated.

After the demonstration and evaluation phase, we can conclude that the framework increases the ability for decision-makers to assess the explicability of machine learning systems. In addition, it creates an overview of the context and supports future improvements regarding the explicability of machine learning systems with specific aspects to improve. Further, it gives an overview of the influential factors for the assessment and creates starting points for future research to ultimately be able to assess the full explicability spectrum of a machine learning system.

Building on that, there has been identified a lot of future research areas that relate to this thesis research. First, the limitations can be coped with by conducting research in the following fields:

- A one-by-one change of explanation characteristic should be investigated on the influence that it has on the EAF, in order to ultimately cover the full explicability spectrum
- Research must be conducted to find the appropriate evidence roles and intelligibility type to include in the framework
- The required threshold should be researched for the soundness and conciseness factor
- Validate the usability of the framework by the actual proposed users
- Empirically investigate the knowledge level of a layperson in the case of credit underwriting

Furthermore, other areas that are recommended to investigate are the following:

- The applicability of the framework in another context that has similarities with the current context, such as automated court decisions
- The scalability of the assessment, since the assessment of multiple explanations at once, increases the added value
- The implication of this research for the automation of explanation generation
- An extension of the expert interviews and the application of the framework on more use cases improves the robustness and validity of the framework even more
- Whether there is a risk of confirmation bias with prospective assessment by the designers of the framework, and what the effect is on the quality of the assessment
- Future research should find out if the lack of seeing the need of incorporating qualitative explicability assessment in the development cycle, 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
- What the costs are for fully 'implementing' explicability in machine learning systems within credit underwriting cases

To finalize, this thesis contributes to the scientific field and the societal field. First, as a scientific contribution, this thesis provides an overview of the interrelations of the influencing factors for the assessment of explicability. Further, it takes a design-oriented perspective and creates a multi-disciplinary, multi-organizational pragmatic assessment framework that guides decision-makers with their choices within the execution of a CRISP-DM lifecycle, in order to create an explicable machine learning system for credit underwriting. The main societal contribution lies in the easier compliance to GDPR and CCD, serving the ethical need of society, and the decrease in unclarity for decision-makers.

Keywords:

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

Table of Contents

List of figures

List of tables

List of key abbreviations

AI	Artificial Intelligence
CCD	Consumer Credit Directive
CoSEM	Complex Systems Engineering and Management
DSRM	Design Science Research Methodology
EAF	Explicability Assessment Framework
GDPR	General Data Protection Regulation
ML	Machine Learning
VSD	Value Sensitive Design

This page is intentionally left blank.

1. Chapter 1. Introduction

1.1 The potential of Artificial Intelligence

One can acknowledge by looking at a (technological) news webpage that Artificial Intelligence is one of the main leading subjects in innovation. This innovative technological field is being thoroughly researched in the current technological transformation age, all across the world. Although Artificial Intelligence (or AI) gained the biggest momentum in the last couple of years, research in this field finds its origins in the 1950s with the *Turing Test* publication (Turing, 1950) and *Dartmouth Summer Research Project on Artificial Intelligence* (McCarthy, Minsky, Rochester, & Shannon, 1955). Whereas the hype at that time caused a lot of interest and investments in research, it was followed by a decline of it in the 1970s, due to high expectations that weren't met, and more realistic expectations of that time gained more support. This cycle repeated itself with the second "AI winter" emerging at the end of the 1980s (Russell & Norvig, 2010). The surge of computing power, the availability of a huge amount of data and the breakthroughs in complex challenges by the use of AI all cause the current interest. The aggregated definition of Artificial Intelligence, perceived from multiple definitions mentioned in *Artificial Intelligence: A Modern Approach* by Russell & Norvig (2010), is as follows:

> *the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment*

The overarching concept of Artificial Intelligence entails many different concepts that can be applied in a broad spectrum of industries and services: Natural Language Processing, Image Recognition, Speech2Text (and Text2Speech), Expert Systems and Machine Learning are just a few of these sub-categories.

The Financial Services Industry (FSI) is eager to apply these techniques in order to become more innovative and utilize more of their available data, as it is the most promising emerging technology for the industry nowadays (Bennink, 2017; Trippi & Turban, 1992). In particular, self-learning algorithms have the potential to reduce costs, improve efficacy, find and create new business ventures and improve risk management (Berthold & Hand, 2007).

We define Machine learning (ML) by combining the definition of Russell & Norvig (2010, p. 4) ("the capability of computers *to adapt to new circumstances and to detect and extrapolate patterns*") and the statement of Samuel (2000, p. 207) ("*programming computers to learn from experience should eventually*

eliminate the need for much of this detailed programming effort"): *the use of self-learning algorithms from experience to adapt to new circumstances and to detect and extrapolate patterns*. The practice of this technique could significantly reduce costs for banks in Know Your Customer (KYC) business cases (Mittal & Gupta, n.d.) and could result in more accurate predictions of the probability of default with credit underwriting (Boillet, 2018; Ince & Aktan, 2009; Lui & Lamb, 2018; ZestFinance, n.d.). Image/voice recognition could improve the way how banks interact with their clients (Partington & Pichler, 2013), for example with more personalized financial advice, and insurance companies could utilize ML algorithms for real-time and more accurate pricing strategies (Balasubramanian, Libarikian, & McElhaney, 2018). Furthermore, machine learning techniques are widely used within high-frequency trading and financial fraud detection (such as credit card fraud) mechanisms. This thesis will focus on the practice of machine learning techniques within the FSI. Shortly, there is a wide variety of valuable applications, but there are also societal risks that should be examined and addressed before deployment should take place these techniques.

1.2 Problem background and research question

The past years have shown that the application of machine learning could result in unpredicted and unwanted outcomes. A problem that has occurred in several scenarios, is the problem that human biases in observed data are being reproduced and even exacerbated by computers with the use of artificial intelligence algorithms (Angwin, Varner, & Tobin, 2017; Crawford, 2016). Two types of harms with bias are identified: 'allocative harm' (e.g. the denial of an opportunity or resources by a system based on bias against a certain population) and 'representational harm' (e.g. the underrepresentation of a certain population which leads to a skewed data distribution) (Crawford, 2017). Both issues are derived from data bias. The main problem with this is that it could result in discrimination and lead to unfair decision-making. Two examples of these unwanted situations are interest-rate discrimination with mortgage lending based on ethnicity (Bartlett, Morse, Stanton, & Wallace, 2017) and a disproportional amount of expensive subprime loans issued to minority groups (Havard, 2011). When this occurs, *accountability* for this problem is very important and this should be a first priority to ensure: if no one can be held accountable for such a problem, we have an *accountability gap* and solving this problem or doing justice is much harder. This is acknowledged as well by the government of the USA ("Algorithmic Accountability Act of 2019", 2019) and the European Parliament (Koene et. al, 2019). One of the essential aspects to ensure this is the *function of explanation* (Dignum, 2017, p. 4703): it should be able to explain an automated decision. Moreover, these

problems need to be avoided. By understanding the underlying workings and the reasoning behind a decision, one can be surer that there is a small probability of such problems.

Another challenge is that black-box models (input and output are known, but not the internal structure (Russell & Norvig, 2010, p.57)) cannot be solely used for automated decision-making, due to the intended right of 'data subjects' to explanation (GDPR, 2016[1]). Despite that the ambiguity of the used language in the General Data Protection Regulation raises questions over the current legal feasibility of such a right (Wachter, Mittelstadt, & Floridi, 2017), *explainable* decisions are ethically desired to understand the decision process. Gilpin et al. (2018, p. 2) state the following on the concept of explainability, or explicability: it should describe *"the internals of a system in a way that is understandable to humans"* and *"the operation of a system in an accurate way".* Moreover, the principle of explicability works towards 'accountable AI', where decisions and actions made by ML-systems "*must be derivable from, and explained by, the decision-making algorithms used*" (Dignum et al., 2018, p. 62).

It is evident that, we are facing a dilemma or at least a tension. On the one hand, we want to maximize the prediction-accuracy and minimize discrimination with machine learning. On the other hand, it is common that the internal processes of a machine learning model become harder to explain when the complexity of the model increases, while aiming to decrease the prediction-error (Sheh & Monteath, 2018). Since the search for explicability could reduce efficiency, cause bias towards explicable but worse-performing models and forced design choices (Adadi & Berrada, 2018; Lipton, 2018), there is an important tension. "*The predictive models that are the most powerful are usually the least interpretable*" (Kuhn & Johnson, 2013, p. 50). In relation to financial services, a problem could occur, for instance, when a requested loan gets declined, and the loan applicant wants to know the justification of this decision. Creditors should inform the consumer about this decision and, moreover, ensure that it is compliant with the right to non-discrimination (EU Consumer Credit Directive[2]) and that this decision is not based on discriminatory factors.

Solely removing sensitive input information is not enough to ensure non-discrimination, and could make it even harder to discover this; discrimination might still happen and the users are not aware of this problem, which even enlarges the problem over time. There are as well more technical problems. *Redlining*

[1] 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), 2016

[2] 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/EE (EU Consumer Credit Directive), 2008

within Data mining could still exist, which means that indirect discrimination is not eliminated while direct discrimination is (Hajian & Domingo-Ferrer, 2013; Pedreschi, Ruggieri, & Turini, 2008); the removal of direct discriminatory factors such as nationality, ethnicity, and sex improves a lot, however, indirect discrimination could still exist through other variables that are still used. Second, the counterintuitive *Simpson's paradox* (Freitas, 2001) could occur; a population split up group A and B seem to have a higher probability of default for group A, however, if both groups are further divided among the similar categories, the probability of default is higher for group B; in short, statistical choices to expose the data can cause wrong conclusions. At last, *Affirmative Action*, or positive discrimination, is sometimes accepted as a means to reduce unequal treatment, but could as well be interpreted as unfairness (Holzer & Neumark, 2006): this does not remove the bias in the trained machine learning model, but counteracts towards it.

Concluding, given the current status of technology, the issues of discrimination risk, black-box model complexity, and technical problems (*Redlining, Simpson's paradox* and *affirmative action)* require that in many cases decisions have to be made regarding an important tension: *explicability versus prediction-accuracy (or performance)*. The ML-systems, as well as the institutional system around it, should be designed in such a way that this tension is systematically taken into account and undesirable ethical, legal and societal effects are minimized by designing the right kind and level of explicability in the systems. This moves towards the idea that ethics can be the source of technological improvement and a reduction of moral overload (van den Hoven, Lokhorst, & Van de Poel, 2012).The evaluation of explicability improves the understanding of the problems that occur here, and with this knowledge, improvements can be made to make a machine learning system more explicable. Herewith, we take the perspective of the current technological landscape. It has become increasingly clear among experts that not just the data science field must be thoroughly researched, but the field of AI ethics as well. Research on how to assess explicability of machine learning applications in the financial services industry creates handles for decision-makers to make choices and support these choices towards the deployment of a machine learning system. Therefore, this thesis aims to answer to following research question:

Main Research Question:

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

1.3 Recent progressions and problem statement

Following the recent advancements and increasing interest of organizations in the ethics of AI, the European Commission appointed an AI expert group to advice (on a high-level) on the decomposition of the general strategy on Artificial Intelligence in Europe (European Commission, 2018). 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.* They are equally important, strengthen each other and should be continuously evaluated, during the design and after the implementation of the AI system (High-Level Expert Group on Artificial Intelligence, 2019). These principles are a good start, but they remain still theoretical and are not pragmatic nor directly usable for evaluation of context-specific cases. The application-specificity of AI-systems requires that high-level principles, as well as low-level requirements, need to be tailored to the case and context (European Commission, 2018; Winfield, 2019).

This thesis 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). There are methods within the machine learning literature that address these problems and enhance the explicability of machine learning systems (e.g. rule extraction and counterfactual explaining). Some of these methods can also be used in credit underwriting cases, but they miss a normative assessment framework in order to substantiate whether a system developed for a certain case (where such a method is pursued) is “explicable enough” to be legitimately deployed. This framework should make the extent to which explanations serve the interest of the different stakeholders of the machine learning system clear and how these explanations enhance the values of these people. The framework is meant for the assessment of use cases, in order to provide starting points to improve the machine learning system on explicability. In addition, the way that assessments need to be implemented in the organizational processes depend on the context of the application case and have to be taken into account. The problem that this thesis tries to solve has been formulated as follows:

Problem statement:

"It is unclear for decision-makers within the development of machine learning systems how to structurally evaluate explicability in credit underwriting cases. To enhance explicability, a pragmatic qualitative framework is needed, useful for prospective assessment of machine learning systems from the point of view of the explicability"

1.4 Knowledge gap

This research is devoted to filling the following knowledge gap that has been identified in the literature overview section (chapter 2.4 describes this more in detail):

The scientific literature lacks research on the prospective assessment of explicability on a pragmatic level (i.e. directly usable for industry practitioners with ML system development) that can guide the development of a machine learning system. Further, it lacks a multi-disciplinary and multi-organizational perspective. Lastly, the literature lacks the aggregation of contextual characteristics of explicability that eventually define the aspects of explicability that need to be assessed.

1.5 Research objective

In order to solve the problem and to fill the knowledge gap the following objective for this thesis is specified, that should satisfy answering the main research question:

"To design a pragmatic prospective assessment framework that can guide decision-makers, within machine learning applications in European credit underwriting cases from the point of view of explicability"

1.6 Research approach

To achieve this objective, the Design Science Research Methodology (DSRM) (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008) has been chosen as the methodology for the design of the framework (figure 2). This is a comprehensive and structured approach that helps to outline the different phases towards the communication of a design (i.e. the assessment framework delivered in thesis format). The objective-centered solution will be used as the entry point as there exists a *"research or industry need that can be addressed by developing an artifact*" (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008, p.14).

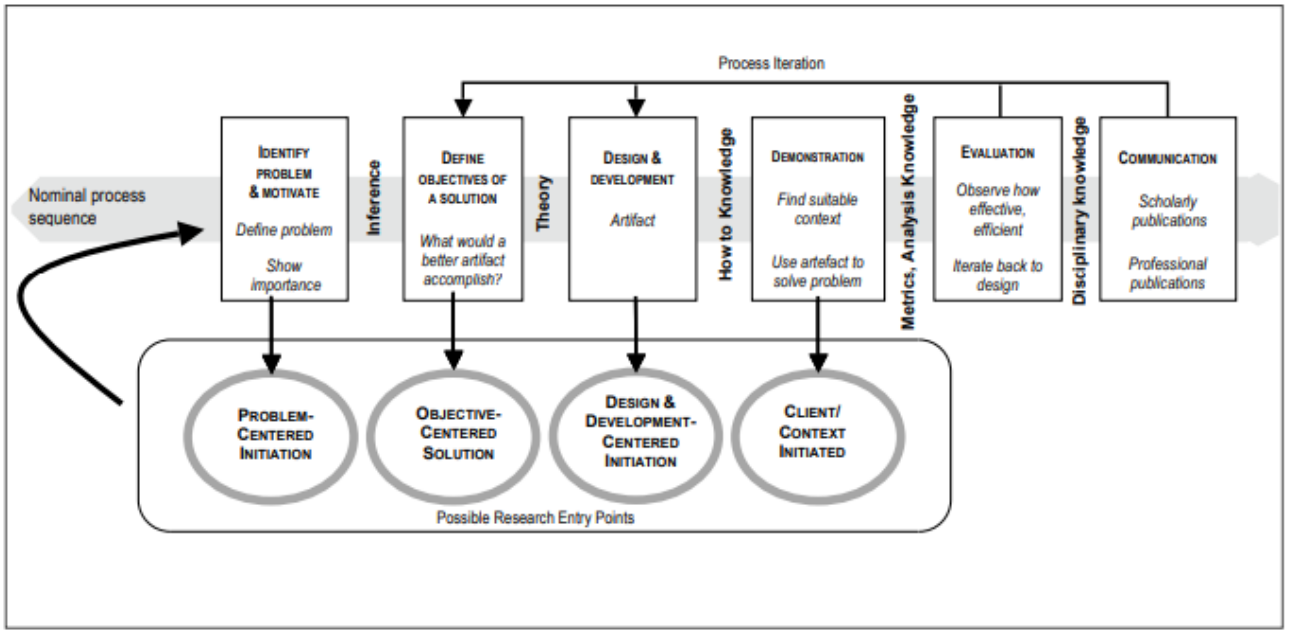


Figure 2: Design Science Research Methodology (DSRM) Process Model (Peffers et al., 2008)

The problem identification (phase 1) is elucidated in the previous sections: the problem statement and the background to motivate this problem are both documented in chapter 1.2 and 1.3. The main advantages of using DSRM are that this method specifically focuses on the design of an artifact (here: an assessment framework), includes a step to evaluate the effectiveness of the designed artifact (pragmatism is included in the objective) and is context-oriented in the demonstration phase (e.g. assessing *explicability* is context-dependent due to the application-specificity of AI-systems).

The DSRM will be complemented with the Value Sensitive Design (VSD) approach (Davis & Nathan, 2015; Friedman, Kahn, Borning, & Huldtgren, 2013). It is defined as: "*a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process"* (Friedman, Kahn, Borning, & Huldtgren, 2013, p. 70). The main advantage of this approach to complement the DSRM is the process of actively incorporating important human values into the design, which is essential to AI-system design. This is as well the reason why the focus is on prospective assessment, opposed to retrospective assessment; before the system is fully deployed the values need to be incorporated in the design, and this can be ensured with the assessment of the design on the value of interest to incorporate (i.e. explicability). By supporting our design methodology with this approach, we make sure that the values of importance in the case of interest are included in the design.

An additional advantage of Value Sensitive Design is its interdisciplinary character because AI is considered *"interdisciplinary per definition"* according to Dignum in an interview (Hansson, 2019).

Chapter 2.5 will elaborate more on the DSRM and VSD approach, and what this means for the development process of the framework of this thesis.

1.7 Master's thesis project relevance

This thesis has as objective to improve the safeguarding of human values, in particular accountability, in machine learning business cases in the FSI, by means of improvement of explicability. It combines technological features as well as these human values. This master's thesis has a twofold of relevance fields: societal and scientific. First, our society will be impacted by ML-applications and it is important that public values are safeguarded during the development of these systems. The Euro area financial sector is highly interconnected (European Central Bank, 2018) and small instabilities could have major effects on trust within this sector. Consumer dissatisfaction, due to bias and unfair AI-applications, may lead to large-scaled commotion and distrust towards the FSI, resulting in an unbalanced financial system. Moreover, article 26 of the International Covenant on Civil and Political Rights (1966), which is part of the Universal Declaration of Human Rights, legally requires non-discrimination among the signatories. Furthermore, human autonomy and transparency are values at stake here. Safeguarding human values in machine learning applications in credit underwriting is therefore of high priority for our society.

Further, scientific relevance is created by means of decreasing the size of the knowledge gap, which is identified in chapter 2. The scientific research field of machine learning ethics requires a way to reduce the gap that exists between the high-level value explicability and the pragmatism that is needed to create explicable machine learning systems. Herewith, a multidisciplinary view has to be taken and the ability to prospectively assess explicability has to be investigated in order to ultimately create a multi-organizational guiding assessment framework. A big priority here is usability of the framework and the alignment with an actual development lifecycle of a machine learning system, such that decision-makers within this development can actually benefit from using it in practice.

1.8 Structure of the thesis

This thesis first continues with a literature overview in chapter 2. Subsequently, the chapters are structured accordingly to the DSRM. Chapter 3 derives the objectives for the artifact from the problem statement and the literature overview. Chapter 4 will go into detail with how the design requirements for the framework

are derived, after which the actual development of the framework can take place. Further, chapter 5 shows the application of the framework on two cases from the scientific literature. Chapter 6 continues with an evaluation of this demonstration and the framework itself on the objectives. The thesis finalizes with a conclusion that includes the answer to the research question, recommendations for future research and the scientific and societal contribution in more detail.

2. Chapter 2. Literature overview

This chapter examines the state-of-the-art literature on the concept of machine learning explicability and close related areas. Currently, there is not a design-oriented literature review present on the subject of machine learning explicability. It is a novel research field which is widely explored now and this thesis takes one path and investigates this from beginning to end: assessment of explicability with a pragmatic design-oriented perspective.

The first part (chapter 2.1) introduces the case of interest in this thesis, which will be used for the demonstration of the framework (chapter 6). In addition, the values of interest for this case sets the first scope for this research. Chapter 2.2 gives an overview of explicability in machine learning, including conceptual clarification of the interrelated values of explicability, transparency, and accountability. Further, it describes the (regulatory and ethical) need for explicability, the people and methods in relation to explicable machine learning. Next, chapter 2.3 discussed different explanation types and evaluation aspects for a good explanation. Chapter 2.4 elaborates more on the scientific knowledge gap to be filled (already described in short in chapter 1.4). Chapter 2.5 discusses the Value Sensitive Design (VSD) approach and the Design Science Research Methodology (DSRM) that will be utilized to design the assessment framework. Chapter 2 finalizes with a summary.

2.1 Machine Learning in the Financial Services Industry

Machine learning is currently being used already or at least researched within the financial services industry, as mentioned in the introduction, with for example high-frequency trading, KYC process optimization, fraud detection, and anti-money laundering investigation. 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, which is the core focus of this thesis. 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.

We scope this further down to personal consumer loans, because the potential fairness issue that the lack of explicability causes is evident here, since a lot of personal data can be used for more precise prediction of credit risk: i.e. there exists *the risk of discrimination (the direct or indirect use of discriminatory factors in the decision-making)*. When discrimination occurs, somebody or some organization should be

accountable for this. Moreover, the design and development process of this system should try to minimize the probability of these issues, and explicability of the ML system is a means towards this minimization. The reason for this is that explicability enhances the understanding of the ML system, and this improves the ability to discover (potential) problems with the system. Furthermore, a stronger accountability relationship is created which increases the possibility to do justice to the accountee (the person who is the subject of the credit decision) in the case of problems.

Within this thesis, a non-exhaustive list of loan type examples is: a loan for financing a car, a house (mortgage), home improvement, a boat, credit card debt or a holiday. It all concerns a human (not a company) who applies at a bank for a certain amount of money, to be paid back in a certain amount of time. Before acceptance of the application, the bank tries to assess the credit risk that it is exposed to if it would accept the application.

This application case of machine learning requires its own case-specific values of interest to be incorporated: accountability, fairness and trust. Trust is lost in AI when users cannot understand the decisions of the systems (Miller, 2019). A bank has an important accountability relation towards their consumers (chapter 2.2.1.3 elaborates more on this) and trust towards the banking industry is required for the stability of the financial system. However, unfairness of decision-making within a bank could cause distrust towards a bank, and therefore from the value-perspective, it needs to be aimed for to minimize this. Thus, the to-be-designed framework will take accountability as 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 Overview of explicability and machine learning

2.2.1 What is explicability in machine learning?

The literature on machine learning (ethics) often refers to concepts such as explicability, transparency, accountability; these are presented as something that should be included with the design of ML systems. However, it is often unclear what the interrelations are between these concepts and which of these are relevant for the concept of an explanation as discussed in this thesis. This chapter will get a grasp on this.

2.2.1.1 Explicability

Explainability and *explicability* are often used interchangeably in the scientific literature. In this thesis, the concept of *explicability* is used, except for the cases that direct citations include the concept of explainable. The reason for this is that the word *explicability* does not linguistically include the *act of explaining*, which

makes it a more neutral word in comparison to *explainability*; it is not dependent 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.

Within philosophy, Lewis (1986) defines explanation as the following: *"to explain an event is to provide some information about its causal history".* Weld & Bansal (2018) extend this with the "*ability to answer what-if questions"*, and the degree that a human can predict how the output will change considering a change of a feature. Gilpin et al. (2018) describe an explicable AI system as a system that can create complete and interpretable explanations (as introduced in chapter 1.2), from the computer science perspective.

Within this context of algorithms and model explanation, Sheh & Monteath (2018) discuss a three-dimensional space from the computer science perspective, for the categorization of explanation techniques: *scope, source,* and *depth.* The *scope* of an explanation can be either *justification* or *teaching*, and since we look at the explanation of a decision, the explanation is called a *justification explanation*. The *source* axis is a continuum between *posthoc rationalization* and *introspective*, and concerns where the information of the explanation comes from. *Introspective source* considers the situation when the decision-making process provides the explanatory information through this same process and "*retains symbolic meaning from that underlying decision*" (Sheh & Monteath, 2018, p. 263). *Post-hoc rationalization source* is the situation where the explanatory information comes from another system (a human) observing the decision-making process, which is the case in the situation we look at; these explanations are merely reflecting on a specific decision. Mittelstadt, Russell & Wachter (2019, p. 2) take a multidisciplinary perspective (researchers in philosophy, machine learning, and law & ethics) and state the following about this: *"Post-hoc human interpretable explanations of models and specific decisions do not seek to reveal how a model functions, but rather how it behaved, and why*". The context of the case of interest defines if such an explanation serves the goal that is aimed for. Lastly, the *depth* of an explanation; i.e. the explanation of a *model* or *attribute*. *Model* refers to information on the training algorithm of the model, the model to be trained itself and background information, and *attribute* to how one (or multiple different) feature(s) of a model influence(s) the final decision.

Within the scope of this paper, the definition of Gilpin et al. (2018) complemented by the definition of Mittelstadt, Russell & Wachter (2019) 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.

2.2.1.2 Transparency

From the machine learning perspective, Weller (2019) argues that transparency in itself shouldn't be a goal, but a means to an end. 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*" (Dignum, 2017, p. 4699). Thus, transparency can be seen as a means to *explicability*. 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.

Transparency is the opposite of the opacity of black-box models (Lipton, 2018). Within the area of model interpretability in computer science, three different perspectives of transparency are distinguished: *simulatability (*model-perspective*), decomposability* (parameter-perspective) and *algorithmic transparency* (training algorithm-perspective). What the knowledge level of the explainee (the person who receives the explanation) is, determines what level of transparency is needed to be able to give a sufficient explanation: e.g. an expert in the field can understand a more technical explanation than a consumer who is not an expert (Weller, 2019).

The goal of transparency ultimately 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 (de Laat, 2018; Weller, 2019). In fact, full transparency to the public could cause issues related to *violations of privacy, undermined efficiency* and *property rights on algorithms* (de Laat, 2018).

2.2.1.3 Accountability

Accountability can be defined in the sense of a social relation (Bovens, 2007 p. 450): *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)" (Dignum, 2017, p. 4703). The function of explanation can be seen as the operationalization of explicability. Therefore, explicability is a means for accountability. 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 (Nissenbaum, 1996).

Accountability with information technology in bureaucratic organizations is often a challenge due to four reasons (Nissenbaum, 1996):

1. Designed and operated by 'many hands'
2. Computer bugs often seem an excuse to not claim accountability
3. It is easy for humans to blame computer instead of the humans involved
4. Software developers often refuse to be held accountable

Next to these reasons, the new situation with more frequent use of opaque models in these IT systems makes the assurance of accountability even a bigger challenge. Doshi-Velez et al. (2017) discuss the context of three categories of tools to ensure accountability with an AI system: *explanation*, *empirical evidence,* and *theoretical guarantees*. The context of the case determines *"why and when explanations are useful enough to outweigh the cost"* and when it can be used as a tool towards accountability. Within this thesis, the explanation category will be investigated.

One can compare the need for accountability in machine learning models with the case of public accountability, as public decision-making has legal social consequences that require explicability as well (i.e. the explication obligation of accountable actors (Bovens, 2007)). Bovens derives seven criteria from his definition that characterize a public accountability relationship (2007, p. 452):

1. *There is a relationship between an actor and a forum*
2. *In which the actor is obliged*
3. *To explain and justify*
4. *His conduct*
5. *The forum can pose questions*
6. *Pass judgement*
7. *And the actor may face consequences*

We can project this 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)

These seven criteria help to capture the institutional dimension of AI explicability and are further elaborated upon in chapter 5. The improvement of explicability simplifies committing to the third characteristic and is, therefore, an important step towards accountability. As can be seen, more aspects should be considered to fully characterize this relationship. The type of accountability we are talking about here is, according to the typology of Bovens (2007), social, hierarchical, legal, ethical and vertical.

2.2.2 The directive, regulatory and ethical drivers for explicability

Additional to the aim to ensure accountability, the Consumer Credit Directive and General Data Protection Regulation are the main drivers from the law perspective for requiring improvement of explicability. Moreover, there is an ethical and societal need for explicability. This chapter will elaborate on these normative drivers for explicability.

2.2.2.1 Consumer Credit Directive (CCD)

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, or the Consumer Credit Directive (CCD), is created to improve the consumer rights in the process of a credit application and agreement with a creditor.

Paragraph 29 of this directive states that a rejection of a credit application, based on data, should be informed by the creditor to the consumer in a way that the fact (of rejection) and the 'particulars of the database' are included. Paragraph 45 states that the directive aims to respect the fundamental rights of the EU, which includes "*the protection of personal data, the right to property, non-discrimination, protection of family and professional life and consumer protection*". Article 9, paragraph 2, specifies this towards database access: *"the creditor must inform the applicant immediately and without charge of the results of such consultation and of particulars of the database consulted".*

Concluding, the consumer has a right to know the decision and what this decision of a credit application is based on.

2.2.2.2 General Data Protection Regulation (GDPR)

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 is more recent regulation that is created for the protection of natural persons regarding the processing of personal data, and has a broader scope than personal credit underwriting, but is projectable on this case.

Five specific paragraphs are particularly interesting in this case. First, article 12, paragraph 1, states that the data controller should take *appropriate measures* to ensure the understanding of information by the data subject whenever it is communicated. It should be in "*concise, transparent, intelligible and easily accessible form, using clear and plain language"*. Paragraph 7 of this article mentions that a meaningful overview of the intended processing should be provided. Article 13, paragraph 2d and f, describes that the data controller should provide the data subject with the information that it has the right to *"lodge a complaint with a supervisory authority"* (in line with characteristic 5 of the public accountability relationship, chapter 2.2.1.3), that automated decision-making or profiling takes place and meaningful information on the logic and effects of this process. Lastly, article 22. GDPR states that "*the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning the data subject and significantly affects him or her"* (paragraph 1). In addition, if in specific situations (it is required for a contract, the data subject explicitly agreed for the decision) paragraph 1 does not apply, the data controller has to "*make suitable measures to safeguard the data subject's rights, freedoms, and legitimate interests*".

Within the scientific community, there is a doubt whether the GDPR is legally binding, concerning the 'right to explanation' (article 22) (Wachter, Mittelstadt, & Floridi, 2017; Bygrave, 2001; Wachter, Mittelstadt, & Russell, 2018). Nonetheless it does show the intentional direction of the European Regulation concerning data, and we should look further than the boundaries of GDPR; individuals have the right to know how their data (and what data) is being processed and this regulatory desire implies the ethical need for explicability, also known as personal- or human autonomy.

Wachter, Mittelstadt, & Russell (2018) conclude that it is better to move beyond the GDPR, due to limitations of the regulation, e.g. the formulation of *"solely on automated processing" (Wachter et al., 2018, p.881),* and I adopt this statement. The adoption of the statement finds its foundation in the fact

that this thesis 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.

2.2.2.3 Ethical- and societal need

Additionally, to these legal drivers, there is the ethical- and societal need in society for improving explicability.

The **ethical need** concerns the higher-level values that the subjects of the decision care about: *personal autonomy, fairness, accountability and explicability*. *Personal autonomy* is undermined whenever a subject of a decision doesn't understand what the decision is based on; the underlying reasoning is missing, and this is needed to exercise autonomy (self-govern their actions) (Buss & Westlund, 2018). Moreover, the subject cannot act on the made decision with thorough reasoning.

Fairness (or non-discrimination) is one of the main principles for the declaration of human rights (see societal need part), and explicability has the ability to help with the assurance of this value. It is based on the principle that "*all discrimination consists of acts, practices, or policies that impose a relative disadvantage on persons based on their membership in a salient social group"* (Altman, 2016) which should be avoided.

Trust is difficult to conceptualize, and this is acknowledged as a complex issue (McLeod, 2015). However, from the practical perspective, it is present when users and subjects of the machine learning system acknowledge that the system does right in good-will. Moreover, *"people must rely on others to act cooperatively and live socially"* (Simpson, 2012), and therefore trust is desired. Taking the pragmatic perspective of this thesis, by creating accountability to these systems with trustworthy institutions, trust towards the machine learning system can be improved, and this enables the cooperation with these systems.

An *accountability* relationship (the leading value for this thesis) is defined in this thesis as follows: t*here is a relationship between an actor and a forum, in which the actor is obliged to explain and justify his conduct. The forum can pose questions, pass judgement and the actor may face consequences" (Bovens, 2007)*. Thus, explicability is needed to enable accountability, and accountability helps in reducing risks and harm (Nissenbaum, 1996).

Explicability ("*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*", as defined in 2.2.1.1) can be seen as a means towards the former mentioned values in this chapter.

Non-commitment to these values could ultimately lead to social rejection of this technology. From the perspective of the company that uses 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 normative institutions put into place to improve ethics. This can take multiple forms, such as developing directives, code of conducts and designing for values (e.g. design the machine learning system with explicability as a value incorporated; the focus of this thesis). The banking code (Ministerie van Financiën, 2015) makes bankers take an oath that lets them *put the interests of customers first* and *to maintain and promotes trust in the financial sector.* The United Nations Human Rights (United Nations, 1948) describe the importance of equality, fairness and non-discrimination in multiple articles. These give additional reasoning for the need for explicability.

2.2.3 People and methods in relation to machine learning explicability

As said, explicability of machine learning systems is dependent on the stakeholders of these systems: different types of explanations may be needed for different actors along the chain in different contexts. This is particularly evident from the pragmatic perspective of explicability. To give an example, a CEO of a company wants to know if a certain system solves a (or multiple) problems, and not the full inner workings, such that it can make a decision on the acceptance of the system.

In addition, there currently are methods that aim to improve the explicability of ML systems. In this chapter, first a commonly used development process (CRISP-DM) is introduced that helps with the understanding of who the different stakeholders are in what phase and how the prospective assessment of explicability could be incorporated in this process. Hereafter, the people that relate to the development lifecycle and machine learning explicability are discussed, after which methods from the literature are introduced that could help with the assurance of explicability of the system.

2.2.3.1 Data science development lifecycle (CRISP-DM)

A machine learning system is being developed according to a chosen development lifecycle. A comprehensive lifecycle model, that has been developed by a consortium of DaimlerChrysler AG, SPSS, NCR and OHRA in 2000, is the Cross-Industry Standard Process for Data Mining (CRISP-DM) (IBM, 2014;

Wirth & Hipp, 2000). It distinguishes 6 different phases that development teams can use as a methodological framework for a data science project: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation* and *Deployment* (Figure 3). This lifecycle concerns the development of a machine learning model.

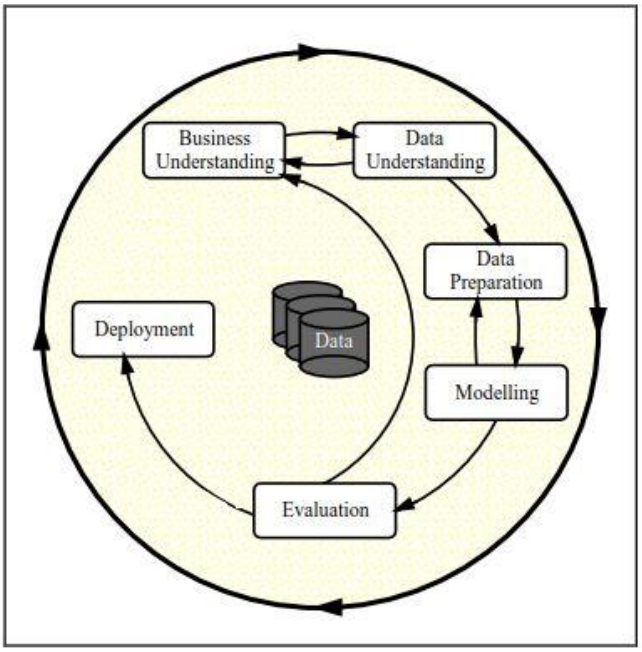


Figure 2: CRISP-DM Process Model (Wirth & Hipp, 2000)

The first phase of CRIPS-DM (business understanding) has the goal to gain insight into what problem the organization is trying to solve on a business level. Project objectives and requirements will be formulated here. To do this, an initial understanding of the data is required and this brings us to the second phase. The close iterative relation between those phases is emphasized with the returning arrow in figure 3. The data preparation phase includes all steps for the dataset transformation, from the raw data towards a more usable dataset. Hereafter, the modeling phase can take place. The close iterative relation means that new problems or ideas often occur when you start with the modeling phase, concerning the data. The evaluation phase finalizes with a decision on whether the model can be used and move to deployment. In this phase, the model and steps towards creating the model should be evaluated based on the business objectives and should evaluate whether the business problem is sufficiently resolved. Explicability needs to be a business objective, so therefore the explicability assessment takes place in this phase. Deployment is the last phase and incorporates a range of possible results, such as a full-scale implementation of a machine learning system, the generation of a report, or a presentation. This concerns a version 1.0 in the

real world and this does not mean that the development lifecycle stops; the development and improvement should be continuous.

One point of critique on this model is, that this process is too abstract from the real world. Nonetheless, using a process model that is more case-specific would cause a less inter-organizational framework and ultimately cuts out generalization possibilities for other cases upfront. In addition, this process is still a norm that many organizations commit to in the real world (Piatesky, 2014), which supports the validity for the real world of using this process model for the development of the assessment framework.

One other point of critique is that it does not include the different perspectives that team-members have; it is a mono-actor perspective. Moreover, this model currently does not enable to directly implement Value Sensitive Design in the process. VSD prescribes the involvement of the different stakeholders in the design process with regards to the validation of the needs of the stakeholders. However, we can observe in figure 4 that there is no such task describes.

Nonetheless, this method will be used with the reason that this mono-actor perspective refers to this current development lifecycle model and the novel addition of explicability assessment can still have a multi-actor perspective, and possibilities to include the VSD. Future research should explore empirical evidence on the "stakeholders' understanding, context and experiences" and value validation.

Foremost the sequence of the steps of this process and the decisions to be made within these steps are useful for the assessment framework. So ultimately this process model will be used to show in what phase explicability needs to be ensured, in relation to the different stakeholders within the development lifecycle and explicability methods, and where these tasks need to be incorporated.

The different phases can be decomposed into tasks and outputs that is visualized in figure 4.

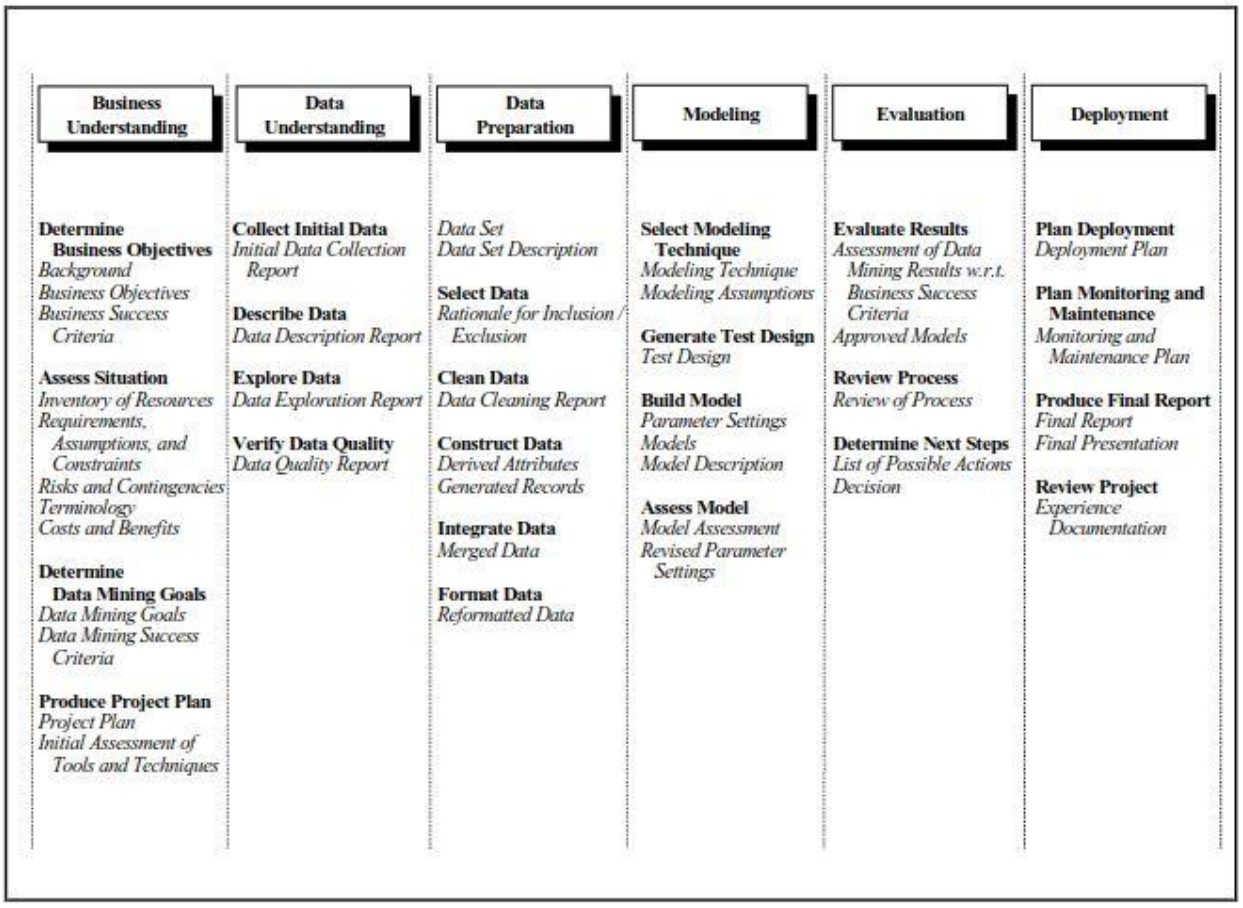


Figure 3: CRISP-DM Tasks and outputs (Wirth & Hipp, 2000)

2.2.3.2 People

Now that we have introduced the process, we can move on to the people that have a certain interest in the case of machine learning explicability assessment, or the stakeholders. We distinguish four main groups and characterize them according to their perceived role, technical competence, responsibility and values/needs (table 1):

Table 1: CRISP-DM Tasks and outputs (Wirth & Hipp, 2000)

Name	Perceived role	Technical competence	Responsibility	Values/needs
Data-scientist	The developers, programmers, and modelers of the machine learning system	Sufficient mathematical and technological knowledge in such a way that they can understand a more technical and mathematical explanation of the machine learning system	Responsible for the design and development of the machine learning system that includes sufficient explicability of the ML system	Explicability towards business employees, laypersons and auditors, good performance of ML system

Business employee	The employees within a company that have knowledge of the business objectives-and requirements of the project	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	They are the ones that are in contact with the consumer if a consumer wants an explanation concerning the decision-making with a loan application.	Accountability with the data-scientists, trust in the data scientists, explicability towards a layperson and auditor
Layperson (consumer)	The loan applicants that want to know why a certain decision was made.	This is a broad group with a wide range of expertise (from no expertise/ knowledge (layperson) to a high level of expertise/ knowledge (expert). Since the group with no expertise nor knowledge is the hardest to satisfy with regards to understandability, the name layperson was given to this group.	The consumer is responsible for valid and truthful data concerning their situation, in order for the bank to be able to assess their credibility	Accessibility and understandability of the explanation, justification of the decision, privacy, fairness
Auditor	The external controllers and validators of the machine learning system after a system is deployed	They are professionals in the field and have a firm understanding of machine learning systems and data governance, and the problems that can occur with them.	They are responsible for checking the compliance of the underlying system to certain principles and standards (that are arguably not defined currently)	Accountability with business employees and data-scientists, transparency, fairness, safety, privacy, security

The groups are formulated in a general way (e.g. no distinction between model developers and validators, positions in the data scientist group), but it is sufficient to describe them like this since we use it for the categorization of explanations and to distinguish the expertise levels. It is not a goal to design a prescriptive task process of who in a company has to do what on a very low-level, however, it is a goal to design a pragmatic framework that can be used by decision-makers that want to assess explicability of the system. Therefore, the framework can be used by multiple roles within the defined groups. An example of this is that a Data Protection Officer (GDPR[3], 2016), who is arguably in either (or both) the data-scientist group or (and) the business employee group, can use the framework to assess explicability. The company who uses the framework can decide better which role specifically can fulfill which task of the assessment in the best way.

The layperson group is the main group that delegates the need for explicability to business employees, since they ask for an explanation on a decision. Furthermore, the auditor group will in the

[3] 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), 2016

future delegate the need for explicability to the data-scientists and the business employees, with regards to principles and standards. Thus, explicability implies different pragmatic relevance to people.

Decision-makers within CRISP-DM

The phases do not necessarily have the same decision-makers. Moreover, different stakeholders define explicability differently (Preece, Harborne, Braines, Tomsett, & Chakraborty, 2018), so this distinction should be made to put this in perspective. Since roles are defined differently within different companies, and the framework is meant to be cross-organizational, a decision-maker will be defined more generally as follows: an employee that is responsible for the final decision that has to be taken within this phase, required to move to the next phase. A decision has pragmatic relevance here. A decision-maker in the lifecycle can be from the data scientist group or the business employee group, as defined in the section before in this chapter (2.2.3.2). Table 2 aligns the decision-makers with the development lifecycle phases, the decisions to be made and the relation to explicability. Since most final decisions in companies are made by the manager at a specific point in time (high in the hierarchy), due to their responsibility, the decision-makers are named as such.

Table 2: Development lifecycle steps, decision-making, and explicability

Development lifecycle phase	Decision-maker	Decide on:	Relation to explicability
1. *Business understanding (Objective formulation)*	Business manager	What objectives and requirements for the business the data science project needs to achieve	Explicability could be a means to achieve one of these objectives
2. *Data understanding (Data governance)*	Data science project manager	Whether the data is useful and of sufficient quality to use in order to achieve the objectives	Explicability starts with at least the understanding of the data
3. *Data preparation*	Data science project manager	Whether the data is ready to be used for model training and testing	Explicability of data can be improved by preparing data
4. *Modeling (Design process of the ML model(s))*	Data science project manager	Whether the quality of the training model and the ML model is sufficient to be evaluated with regards to the objectives	A need for explicability has an influence on what level of complexity of models can be used
5. *Evaluation (Decision process of the ML model)*	Data science project manager	What training and what ML model is the best to use, and if it achieves the objectives that have been formulated	Assess the model(s) on explicability
6. *Deployment (Model implementation)*	Business manager	Whether the model solves the problem of the business with fresh real-world data	Explicability should be aligned with the way how the deployment phase is executed

Within most banking institutions the model validation employees and model development employees are separated, which both fall into the same group 'data scientists' regarding the earlier categorization, due to their same end-goal of the machine learning system and the same expertise level. The assessment framework will be developed for the model developers (that have a manager position), in line with the Value Sensitive Design approach; the design of the machine learning system must incorporate the value *explicability*, and the assessment can help them to evaluate if a sufficient level is reached. Afterwards, the model validators can assess it for a second check with another perspective.

Notice here that auditors and consumers are not included in the table since they are not involved in the development lifecycle. However, they should be included in the considerations regarding the assessment of explicability towards them; i.e. is the system explicable enough towards the auditors and the consumers after deployment?

2.2.3.3 Methods

There are currently multiple methods in the literature that try to make machine learning models more explicable. They reach from more technical/mathematical methods towards methods that try to reduce the complexity of models and result in understandable and human-interpretable explanations of (certain aspects of) the machine learning model. This paragraph will elaborate on these methods

Techniques have been developed that in some specific context can help to provide an understanding of the decisions that a model makes. Frosst & Hinton (2017) use a soft decision tree (trained with stochastic gradient descent) to improve the explicability of a particular decision of a neural network. A variety of methods to improve the explicability of models is addressed by Gunning (2017): *Deep Explanation, Interpretable Models (such as decision trees), Model Induction, Local Interpretable Model-agnostic Explanations (LIME)* (Ribeiro, Singh, & Guestrin, 2016)*, SP-LIME and SHapley Additive exPlanations (SHAP)* (Lundberg & Lee, 2017).

All these techniques are focused on technical explanations of either decisions or the system workings and therefore are not directly pragmatic for explanations towards consumers, i.e. the explanations provided by these techniques should be made more understandable by other techniques.

One method to make a black box model more explicable is by means of rule extraction (Setiono & Liu, 1995). Using backpropagation, the authors extract rules from the neural network to improve the understandability while looking at a minimum amount of conditions (more rules cause more hinder for humans). This minimum amount of conditions goal is important to ensure the human-understandability, as a large number of rules increases the understandability.

In addition to these more technological methods regarding explicability, the literature on robotic task planning is explored in more detail, since explicability has been investigated as a more practical problem here. An annotation has to be made that this research area has characteristics that the research field of machine learning explicability does not have, such as a *physical observation* aspect (an observer can literally see a step that the robot makes, as opposed to a machine learning algorithm step that one cannot observe in the physical world). However, there are useful concepts and ideas in this area that have to be considered for machine learning explicability, to move from the theoretical field (philosophy, ethics) towards more pragmatism of operationalizing this, without turning directly to the technological field.

In the field of robotic task planning, a more explicable model is seen as the model that comes closer to the observer's (explainee) model (Tathagata Chakraborti, Kulkarni, Sreedharan, Smith, & Kambhampati, 2018; Sarath Sreedharan, Chakraborti, Muise, & Kambhampati, 2019; Sreedharan et al., 2017).

T. Chakraborti, Sreedharan, Zhang, & Kambhampati (2017) acknowledge the importance of the interaction with humans in their paper and use the description of a *minimally complete explanation, or MCE, ("shortest model update so that a given plan is optimal in the robot model is also optimal in the updated human model"* (Tathagata Chakraborti et al., 2017)) to formulate the explicability requirement as the ability to generate plans that make sense for a human being. Further, the generated MCE should achieve a certain level of *completeness* and *conciseness* (Tathagata Chakraborti et al., 2017). This relates to what Williams, Szafir, Chakraborti, & Ben Amor (2018) say about explicability in human-robot interaction; the ability for humans to read and recognize the robot's actions enhances the interpretability of the model. Zhang et al. (2017) mention that explicability entails the easiness of the association of tasks with actions for humans. Within the domain of robotic task planning, they make a distinction between the mental model and the model of a robot. The model reconciliation process, as mentioned by S. Sreedharan, Chakraborti, & Kambhampati (2018), could aim for the explanation to many explainees or an explanation where the model of the explainee is unknown. The paper of Sengupta, Chakraborti, Sreedharan, Vadlamudi, & Kambhampati (2017) discusses the distinction of the human and robot model within the case of decision support by robots. Hereby the human should be able to understand the rationale of the decision by the robot model. The level of importance of this understanding relates to the degree of automation of decision-making (figure 5).

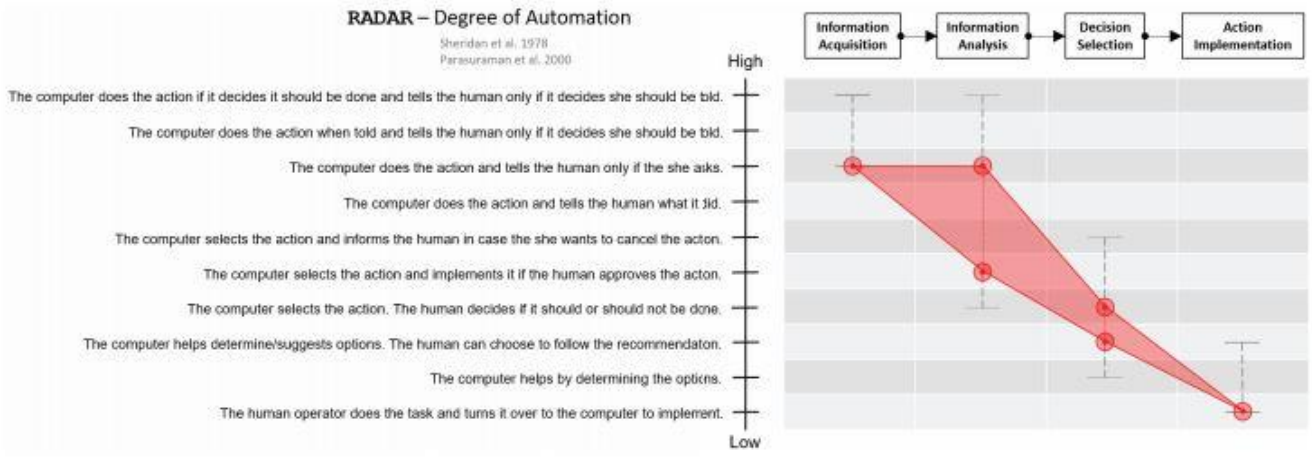


Figure 4: Degree of Automation (Sengupta et al., 2017)

Taking a look at this figure, we observe that in the fourth degree from the bottom ("The computer selects the action. The human decides if it should or should not be done.") a decision is made automatically by the computer. Although in the first degrees on a human still decides if it adopts the decision, this decision needs an explanation. Moreover, from the fourth degree on explicability of the machine learning system is a requirement for ethical use of the system.

Other influential work that is focused on explainable artificial intelligence (XAI) is that of Gunning (2017) and the Defense Advanced Research Projects Agency (DARPA). Gunning (2017) provides a high-level explanation model (figure 6) that includes concepts such as *Explanation*, the earlier mentioned *User's Mental Model* and *Trust* as one of the drivers for the need of explanations, like in our case of interest. They've developed an XAI System One can see that the assessment of an explanation is essential in order to sustain the quality to update the *User's Mental Model* for the better and eventually improve *Trust*.

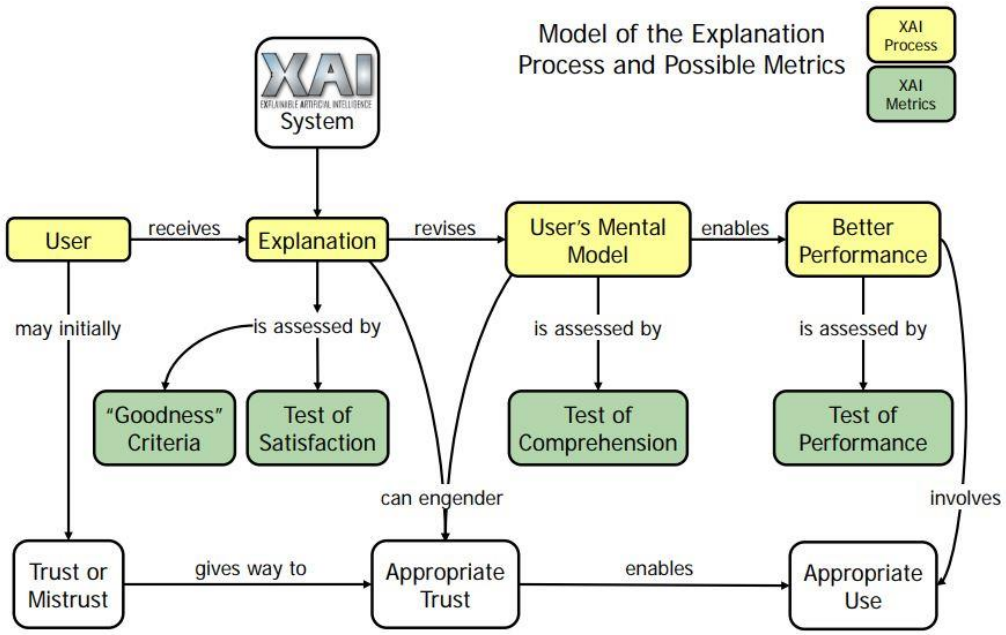


Figure 5: Psychological Model of Explanation (Gunning, 2017)

Miller (2019) argues that within the artificial intelligence community the goal of explaining is often to create a shared understanding of the decision to a certain level. This goal is shared with Kulkarni et al. (2019), and with Lombrozo (2006, 2012) who sees explanations as a way to bring new information to prior beliefs and moves towards generalization.

This paper adopts the statement made regarding *contrastive explanations* by Miller (2019): people want to know the explanation to understand what the cause of a decision is, relative to some other decision that did not occur (the outlier). This idea is comparable to the concept of updating the mental model of the user through an explanation, when there is a difference, as we've seen in the literature on robotic task planning. The reason for this adoption is the focus on *explicability as an instrument for the value of accountability in machine learning* that is central within this paper; accountability is especially desired in situations where an expected decision did not occur (e.g. a loan application was not accepted). Borgo, Cashmore, & Magazzeni (2018) acknowledge the idea of contrastive explanations and apply it to a methodology for explainable AI within AI planner decisions. This is based on the idea that users will suggest different actions that they expect over the action that is in the plan and will ask about the explanation behind this different action.

Fox, Long, & Magazzeni (2017, p. 2) provide 6 questions in the research field of robotic planners that "*characterize what it means for the behavior of a planner to be explainable"* that are used by the methodology of Borgo et al. (2018) as well:

1. *Why did you do that?*
2. *Why didn't you do something else?*
3. *Why is what you propose to do more efficient/safe/cheap than something else?*
4. *Why can't you do that?*
5. *Why do I need to replan at this point?*
6. *Why do I not need to replan at this point?*

The fifth and sixth question are context-specific for robotic planners, and therefore not applicable to our case. The first four questions can be transformed in questions to be asked to an explainer of a machine learning system that makes a decision on a loan applicant:

1. Why did you make this decision?
2. Why didn't you make another decision?
3. Why is this decision better than another decision?
4. Why can't you make this decision?

Answering these questions is expected to provide insights into the reasoning for a certain decision and the system. One can see that the method of answering these questions to provide explanations does not imply any specific level of abstraction of the explanation, and thus this method does not imply a required level of expertise to be the explainer or the explainee. However, the explainer that uses this method needs to ensure that the explanation is in line with the level of expertise of the explainee.

2.3 What is a good explanation?

These explanation methods of chapter 2.2.3.3. ultimately produce an explanation. This chapter first outlines the different types of explanations that are out there and defines the scope of the assessment framework regarding the type of explanation to be assessed. Further, it discusses the goodness evaluation of, which will be used to form objectives for the framework in chapter 3.

2.3.1 Explanation types

Mueller et al. (2019) use a wide variety of literature from research in philosophy, cognitive science, human-computer interaction and psychology to explore what is recognized as a 'good' explanation, which is a main

topic currently in the research field of AI ethics. A firm understanding of what an explanation is, needs to be formed (Guidotti et al., 2018). The first main question to be answered is "What is explanation?" (Miller, 2019, p.11) and it can be categorized in trifold:

- Explanation is 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 is a product: "*the explanation that results from this cognitive process is the product"* (e.g. a textual, visual or conversational explanation)
- Explanation is 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 7). 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. Furthermore, this product that needs to be transferred to an explainee defines ultimately how the social process needs to be designed and assessed. The cognitive process 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.

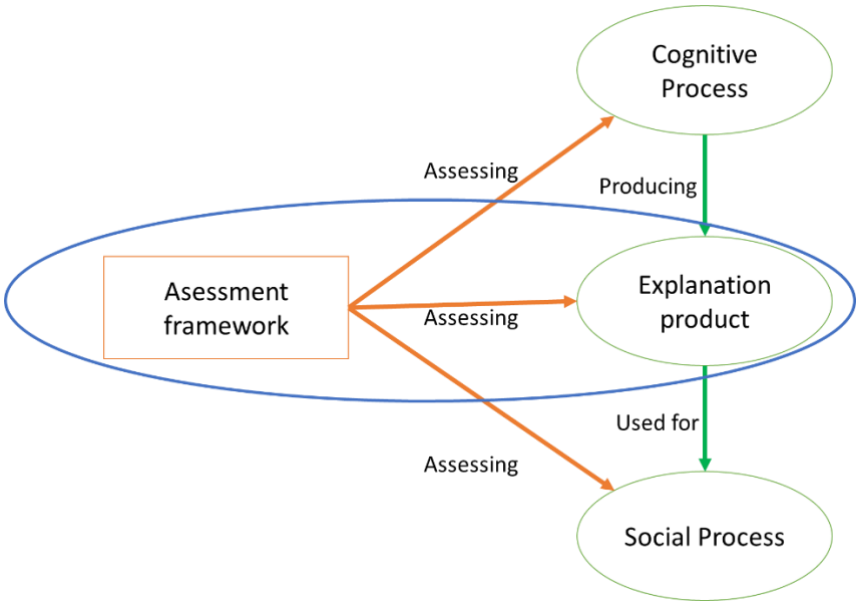


Figure 6: Assessing the explicability process

2.3.1.1 Explanation as a cognitive process

This *explanation* has already been discussed in chapter 2.2.3.3 and is on the processes that eventually formulate an explanation (product). Hereby the question that needs to be answered (in future research) 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 (Guidotti et al., 2018, p. 12-15). Examples of recent comprehensive methods within the classes are given underneath the description of the class, extracted from the papers of (Guidotti et al., 2018; Cui, Lee, & Hsieh, 2019; Gunning, 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017; Gilpin et al., 2018):

- *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. by use of (automatic) rule extraction from trained deep neural networks
- *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. by use of the Local Interpretable Model-agnostic Explanations (LIME) approach (Linear Proxy model), Model Induction, SP-LIME
- *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. by use of a sensitivity analysis, salience mapping, counterfactual XAI technique, convolutional neural network, feature importance ranking measure, deep explanation, Random Forest, SHapley Additive exPlanations (SHAP)
- *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 do not require another model to mimic the behavior, in order to provide locally or globally human-interpretable explanations
    - e.g. the design of a decision tree classifier model, the design of a linear regression model

Miller (2019) argues that asking 'why-questions' (why, why not, what-if) to systems improves the understanding, is more efficient than providing the full causal explanation, and thus can be a means to produce explanations. This method, that can be used in the situation of ex-post decision investigation, is especially of interest in the context where a consumer expects a decision and this decision did not occur (contrastive towards the latter). An example of the form that an answer to these questions can take is the form of a counterfactual explanation (Wachter et al., 2018), which will be discussed in chapter 2.3.1.3.

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 Explanation as a social process

Most literature on methods for explanation is from the technical perspective and they do not incorporate the social aspects of explanation: transferring the formulated (proposed understandable) explanations to the explainee and the social interaction between the explainer and the explainee in relation to this.

Alignment with societal values and ensuring the understanding of public opinion is very important here, as mentioned by Floridi et al. (2018). This implies that validation of these aspects with the consumer should be included in the social process of explanation and an evaluation feedback loop possibly helps to achieve this. Further, one needs to validate whether the explanation is understandable for the explainee, as this differs per individual, and some individuals need more elaboration or more extensive explanations to achieve the same level of understandability. Doshi-Velez & Kim (2017) argue for three different approaches of evaluations of explanations that can be chosen to accomplish this: *application-grounded, human-grounded* and *functionally-grounded,* where the first type is the most specific and the last type the least.

Herewith the level of automation of this feedback loop is another aspect that needs consideration since there are multiple ways that the feedback loop can be designed and implemented.

Considering the context of the credit underwriting case, the application-grounded evaluation seems the best fit, as there exists a concrete application. On the other hand, this is be a long-lasting process. To shorten this, a larger target community can be reached with the adoption of the human-grounded approach. An optimal mix of multiple approaches should be aimed for, that depends on the case and company resources.

The context of the case defines also the goal of transferring the explanation which is either justification or teaching the explainee (Sheh & Monteath, 2018). Relating this question to the value of accountability and the public accountability relationship (Bovens, 2007) of the case (chapter 2.2.1.3), it is concluded that the goal is not to learn a consumer something, but **explaining with the goal to justify a decision.** This case-specific explanation goal that is adopted, or explanatory value, affects the type of explanation that will be assessed as well, and this will further be discussed in the chapter on explanation as a product (2.3.1.3).

It is interesting to look at the literature in healthcare with regards to explicablity; accountability is a highly required value within healthcare. This medical field is thus another area where explicability, or the "*ability to explain decisions accurately"* (Papageorgiou et al., 2006, p. 67), is important in order to understand the causality of an observed output. Mistakes in the method can be very harmful. A high level of explicability enhances trust in the method (Samek, Wiegand, & Müller, 2017), which is also needed in some applications of complex machine learning models, such as credit underwriting. The medical domain emphasizes that machine output should be *human-verifiable* and *precise* to support decision-making (Holzinger, Biemann, Pattichis, & Kell, 2017). Social studies area should eventually investigate the possibilities for this social process and find out what works the best in this justificatory context.

2.3.1.3 Explanation as a product

In order to create explicability, the explanation *utterance* needs an "*unmistakable form – at least with respect to the given situation"* (Kannetzky, 2002). The consumers' demand and situation indicate what form the explanation needs to take, what aspects to address and at what moment, because a full investigation could be too excessive. Lim & Dey (2009) recommend a list of design implications for intelligibility. These different types of intelligibility are: *application, situation, input, output, model, why, why not, how, what if, what else, visualization, certainty, and control.* The context of the case in scope defines ultimately which intelligibility types should be explained (Doshi-Velez et al., 2017). The intelligibility

types that are required ultimately lead to a certain explanation producing method, or multiple methods, to be implemented.

To assess the justificatory aspects of an explanation (justification is the goal of the explanation in scope) we make use of the evidence roles, as defined by Biran & McKeown (2014). It concerns the following evidence roles:

- 'normal evidence'
    - E.g. evidence expected to be present in many instances predicted to be in this class (high positive importance, high positive effect on the prediction)
- 'exceptional evidence'
    - E.g. evidence that is not usually expected to be present (low importance, high positive effect on the prediction)
- 'contrarian evidence'
    - E.g. strongly unexpected evidence, since the effect has the opposite sign than expected (high negative importance, high positive effect on the prediction)
- 'missing evidence'?
    - E.g. important features that were expected to contribute highly positively on the prediction, but were weak for the prediction (high positive importance, low effect on the prediction)
- 'normal counter-evidence'
    - E.g. expected to contribute negatively on the prediction and does contribute negatively on the prediction (high negative importance, high negative effect on the prediction)
- 'exceptional counter-evidence'
    - E.g. unexpected, the feature is not expected to contribute highly negative on the prediction, but it does contribute highly negative (low importance, high negative effect on the prediction)
- 'contrarian counter-evidence'
    - E.g. feature we expect to contribute positively, but contributes negatively instead (high positive importance, high negative effect on the prediction)
- 'missing counter-evidence'

- E.g. feature that was expected to contribute highly negatively, but was weak for the prediction (high negative importance, low effect on the prediction)

These evidence roles need to be chosen in the framework application such that a decision is justified by these evidence roles. Further, they need to be chosen in such a way that unexpected evidence is included in the explanation if this exists. This has pragmatic relevance, because these evidence roles could imply issues. Moreover, the reason why such evidence exists needs to be investigated by data-scientists. If such a situation exists in the case that a layperson gets an explanation with unexpected evidence, he or she has information for recourse; he or she can act upon it. It has pragmatic relevance on one other side as well; when a loan gets declined the applicant can try to change the reason for it such that the credit application gets accepted in the future.

Explanations as a product can be categorized in trifold, based on the *moment in time* that concerns the decision and the *scope* of the explanation (Wachter et al., 2017): *ex-ante/ex-post* (prior to an automated decision and after an automated decision) and *global/local* (system functionality/specific decision).

- *Global ex-ante:* explanations of the system functionality before the decision is made
- *Global ex-post*: explanations of the system functionality after the decision is made
- *Local ex-post:* explanations of specific decisions that are made by the machine learning system[4] (notice here that this can include hypothetical decisions as well, to investigate a specific scenario).

Local ex-post explanation will be the scope within this thesis, as this type is the most relevant for the GDPR and CCD (justificatory value is required), for ensuring the explicability of this type is in line with the fourth characteristic of the public accountability relationship (chapter 2.2.1.3) and since accountability is the upper value of interest.

In addition to these three types of explanations, it needs to be evaluated which *level(s) of abstraction of explanation(s)* is (are) required in what context (Seegebarth, Müller, Schattenberg, & Biundo-Stephan, 2012). What level is needed, depends on who the recipient of the explanation is (e.g. what is their experience with the application domain, singular or a group of people), and what the goal of the explanation is. Examples of this are to make a decision understandable for a layperson, to monitor whether a system is logically sound within the business goal and process, or to check if the mathematical logic within

[4] Ex ante local explanations do by definition not exist, due to the fact that a specific decision in a certain context cannot be explained *before* the decision is made

the machine learning model is correct and no mistakes are made. An assumption is made here that the layperson has English as a native language since the explanations will be in English. Other languages as native language could give additional unnecessary complexity since the outcomes of the research are prospected to be easily adapted to other languages.

Within this thesis, three definitions of different abstraction levels are distinguished and will be applied: *machine-level, business-level,* and *consumer-level* (figure 8). This can be related to interpretability: Doshi-Velez & Kim (2017) define interpretability as the "*ability to explain in understandable terms to a human"*. We extend this definition to two types of interpretability: machine-interpretable and human-interpretable, where the first refers to the assumption that the explainee (i.e. the human in the definition) has a more advanced knowledge level of data science, so that it can understand explanations regarding mathematics, statistics, and computer science where required. The latter refers to the assumption that no further knowledge is required other than basic human language understanding. This has as a result that machine-level explanations have *explanatory value* for data scientists, but not for laypeople.

- *Machine-level* concerns a mathematically sound explanation so that a data scientist and/or mathematician can validate the quality of the algorithms, calculations and model outcomes; i.e. the bottom abstraction level. 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. This is the most machine-interpretable level.
- *Business-level* relates to 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; i.e. the middle abstraction level. 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* 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; i.e. the top abstraction level. 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). This is the most human-interpretable level.

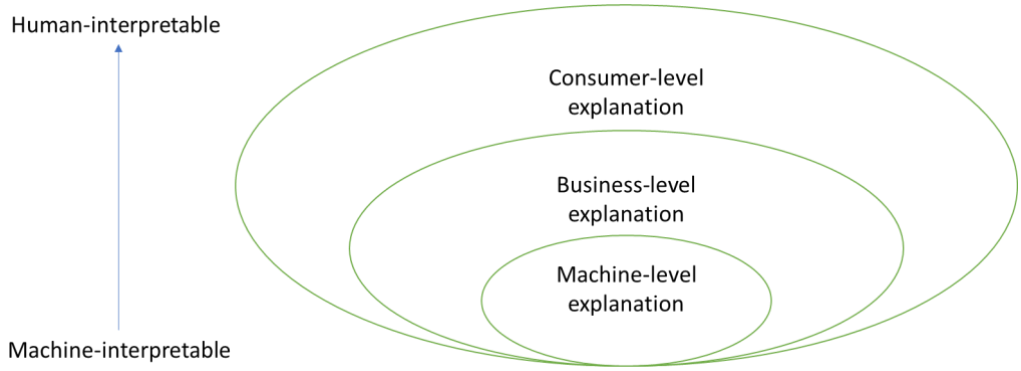


Figure 7: Machine-, Business-, Consumer-level explanation levels of abstraction

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.

We can compare the description of a machine-level explanation to the definition of *transparency* by Dignum (2017, p. 4699): 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*", or 'opening the black-box'. Further, machine-level explanation takes transparency to a more pragmatic level, as this includes as well the process of formulating an explanation based on these internal workings, taking a technical and mathematical form, and the social process of transferring an explanation to an explainee.

Concerning the GDPR, the consumer-level explanations are the most of interest and this is the scope of the thesis. 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 for an explanation is the most interesting, since the consumer-level has the most to offer for them. Wachter et al. (2018) argue that *counterfactual explanations* is a means to explain specific decisions (local ex-post). The counterfactuals take the following form (Wachter et al., 2018 p. 848): "*Score p was returned because variables V had values (v1,v2,…) associated with them. If V instead had values (v1',v2',…), and all other variables had remained constant, score p' would have been returned*". They do not require to open the black-box to provide an explanation that can help to inform individuals, in order to comply with future legal rights as intended by the GDPR. Opposed to this, these explanations are not sufficient in the scenarios where an understanding

of internal workings, the rationale of a decision or statistical evidence is required (Wachter et al., 2018). Thus, it can provide reasons why a particular decision has been formed, and this could enhance understanding of the decision and create a base for individuals to question (and fight) a decision, however, it cannot provide the (statistical) evidence that discrimination did **not** occur at all. Besides that, this technique is able to provide evidence that a decision **has** been influenced by the use of a discriminatory variable if this is the case.

2.3.1.4 'Explanation type' scope

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 3. The chapters in which the choice for the specific characteristic is discussed, are mentioned next to the characteristic. 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. The assessment ability only exists with the logical combinations.

Table 3: Explanation assessment types

Class name	**Characteristics of the explanation of interest**	**All possible characteristics of the class**
Assessment moment	Prospective (chapter 1.6)	Retrospective, prospective
Explicability process sub-part	Explanation product (chapter 2.3.1)	Cognitive process, social process, explanation product
Explanatory value	Justification (chapter 2.3.1.2)	Justification, teaching
Explanation scope	Local (chapter 2.3.1.3)	Local, global
Moment in time	Ex-post (chapter 2.3.1.3)	Ex-post, ex-ante
Level of abstraction	Consumer-level (chapter 2.3.1.3)	Machine-level, business-level, consumer-level
Explainee	Layperson (chapter 2.3.1.3)	Data-scientist, business, auditor, layperson (consumer)

2.3.2 Explanation goodness evaluation

The definition of "good" metrics for explanation is still an important issue currently, according to Fox, Long, & Magazzeni (2017, p. 2). Gunning (2017) addresses measures of Explanation Effectiveness,

however, they are not sufficiently pragmatic for decision-makers within machine learning cases to use. Seegebarth, Müller, Schattenberg, & Biundo-Stephan (2012) label four other factors that can improve the quality of explanations: *the length of an explanation, level of abstraction, the fitness to context and the referenced steps from the model or plan that is being explained.* They suggest as well that future research should be focused on quality measures of explanations within a certain application context and the ability to give explanations in an informative human-readable way or research on the outcome explanation problem as defined by Guidotti et al. (2018).

Hoffman, Mueller, Klein & Litman (2018) provide a list of features that are intended to facilitate a "good" explanation: *understandability, satisfying, detailed, complete, actionable, accuracy/reliability and trustworthy.* However, these features are still subjective and susceptible to wrong interpretation. Mueller et al. (2019) list *soundness, appropriate detail, veridicality, usefulness, clarity, completeness, observability, and dimensions of variation* as the features that a good explanation should possess. Cui, Lee, & Hsieh (2019) evaluate XAI techniques on 3 criteria: *Correlation, Completeness, and Complexity*. They stated that they've filtered out multiple criteria from the earlier lists due to the fact that they solely focus on functional explanations (closing the information gap).

Kulesza, Burnett, Wong, & Stumpf (2015) suggest four principles for good explanations, that will be adopted. It summarizes principles or criteria mentioned in other papers and transfers it towards a comprehensive list. Adopting these criteria in explanations should increase the understanding of the machine learning system (Kulesza et al., 2013): *iterativeness (or conciseness), soundness, completeness, and non-overwhelming* (*or comprehensible).*

- Iterativeness or conciseness: "*concise, easily consumable bites of information, that users can attend to if interested*" (Kulesza et al., 2015, p. 2)
- Soundness: "*the extent to which each component of an explanation's content is truthful in describing the underlying system*" (Kulesza et al., 2013, p. 2)
- Completeness: "*the extent to which all of the underlying system is described by the explanation*" (Kulesza et al., 2013, p. 2)
- Non-overwhelming or "*comprehensible";* selective in features to explain (Kulesza et al., 2015, p. 3)

Despite that conciseness of an explanation will be accepted, the iterativeness will be neglected because this aspect better fits the social process of explanation. Comprehensibility is extended with the linguistic aspects, such as choice of words. The characteristics of soundness, completeness, conciseness (defined as 'compactness' by Guidotti et al.) and comprehensibility are also acknowledged by Guidotti et

al. (2018) as important goals to aim for, with a remark that the context properties determine the level of importance of these characteristics such as “the expertise of the user” (Guidotti et al., 2018, p.36)).

To ensure the quality of an explanation, an explanation should possess these four characteristics as defined (Kulesza et al., 2015, 2013). Two experts acknowledge in an interview (Appendix 2) that the type of explanation that is being assessed defines to what extent these characteristics have importance for the explanations:

- *Completeness*: very useful, but completeness does not imply directly more trust towards a system. This characteristic is more important for global explanations.
- *Conciseness:* always important, but has tension with completeness. An optimal point between those characteristics should be aimed for.
- *Soundness*: essential, this should be aimed for within all types of explanations
- *Comprehensibility*: an important aspiration, however, the question here is to what extent this should be accomplished. 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.

The framework should make it possible to assess the explanation product on these four principles. 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.

On the other hand, if an explanation is sufficient to fulfill the explicability requirement for consumers, the machine-level and business-level explainers should ask their selves if it is possible to make the model less complex; i.e. why would you deploy a model that is very complex, that can eventually be explained by a human, and might be able to be replaced by a simpler model? (Robbins, 2019) Unnecessary complexity should be avoided. Besides that, it should be borne in mind here that the final goal of the use of a more complex model is to improve the prediction-accuracy, and that on the other hand, the change towards a simpler model *could possibly* result in a reduction of prediction-accuracy (again: *the explicability-performance tension*). Therefore, a good strategy is to start at the beginning of the development lifecycle with simpler models and move towards more complex models if these significantly improves the performance.

2.4 Knowledge gap (in-depth)

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.

2.5 Methodology: Value Sensitive Design (VSD) and Design Science Research Methodology (DSRM)

The methodology that will be utilized within this thesis is the Value Sensitive Design approach in combination with the Design Science Research Methodology. First, the VSD approach will be explained, after which the more guiding methodology DSRM will be elaborated upon. Further, the research scope will be discussed, and a visual representation of the research methodology will be given by the use of a '*research flow diagram*'.

2.5.1 VSD approach

Value Sensitive Design, 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 (Davis & Nathan, 2015). Within this thesis, 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 ("*Ideally, Value Sensitive Design will work in concert with organizational objectives."* (Friedman et al., 2013, p. 77)*)*

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.

Davis & Nathan (2015) discuss the tripartite methodology (figure 9) which in short entails distinguishing the following three interdisciplinary investigations:

- Conceptual: identifying stakeholders and define values implicated by the use of machine learning
- Empirical: examine stakeholders' "understandings, context, and experiences" (value validation)
- Technical: the study of specific features of technologies and values mapped to design-requirements

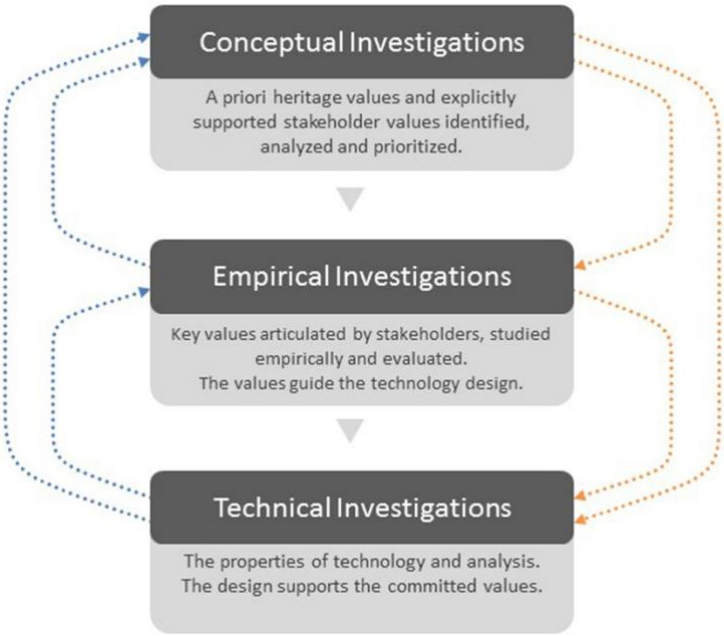


Figure 8: Value Sensitive Design investigations (Mok & Hyysalo, 2018)

The importance of these three investigations is context-specific (Davis & Nathan, 2015). For our case of interest, the conceptual investigation is the most important; the identification of the stakeholders, their values, and the guiding value for which the framework will be designed. This has already been accomplished in chapter 2.2.1 and 2.2.3.2.

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 with potential stakeholders to validate the objectives (appendix 2). Real-world use of the framework should carry out a supported conclusion regarding the validity. A possibility to accomplish this is through including this validation in the *social process* of *explicability* as documented in chapter 2.3.1.2.

As mentioned earlier (chapter 2.2.3.1), VSD asks for inclusion of stakeholders in the process. Since the current CRISP-DM not really incorporates this aspect, the new addition of explicability assessment is a way to push this method more towards the Value Sensitive Design approach. Understanding the values of the stakeholders is one major step in the right direction and including this as requirements in the development cycle is a way to create interdisciplinarity and involvement of stakeholders (which helps in creating trust in the system).

2.5.2 DSRM

The DSRM by Peffers, Tuunanen, Rothenberger, & Chatterjee (2008) 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) (Taebi, Correljé, Cuppen, Dignum, & Pesch, 2014), the Design Science Research Methodology (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 (Vaishnavi, Kuechler, & Petter, 2017) like it will be in this thesis. The other output is a scholarly or professional publication of the obtained disciplinary knowledge, which will be in the form of a graduation thesis (and a scientific paper). Using this qualitative methodology has the advantages of literature-based, practical guidance and a model that provides in a presentation (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008). 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 that the sub-questions are linked to, and chapter 2.5.2.1 will discuss them. Since the framework (the to-be-developed artifact) aims to solve the research- and industry needs to be able to prospectively assess explicability of a machine learning system, this research enters the DSRM approach at the objective-centered solution entry point.

The artifact that will be designed has the following characteristics:

- A framework for prospective assessment of explanation as a product
    - It concerns a justificatory local ex-post explanation on consumer-level towards a layperson
- It is pragmatic for decision-makers within the development lifecycle of machine learning systems to use for guidance with decisions concerning the question if a certain system is explicable enough to be implemented
- It is adjusted towards the context of the credit underwriting case for personal loans in banking
- It is adjusted towards the values of *accountability* and *explicability*

2.5.2.1 Research phases and sub-questions

To answer the main research question by using the DSRM, six sub-questions (SQ) are formulated that are linked to the different phases of this methodology; the sub-questions need to be answered.

Phase 1: Problem identification and motivation. (Chapter 1)

[SQ1]: What is the problem that the objectives should solve?

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. This phase has already been pursued in chapter 1.1 to 1.3. The objectives of the framework should at least be formulated in such a way that this problem can be effectively solved by the framework. Literature research has been performed in order to investigate and motivate the problem, and ultimately formulate a problem statement.

Phase 2: Objectives of the solution. (Chapter 3)

[SQ2]: What are the objectives for the assessment framework?

The next 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) in order to validate if the framework has solved the earlier stated problem. The research aims to formulate assessment factors that are applicable for the specific case and explanation type of interest, but might be generalizable to another context as well (which will be discussed in the communication phase). Further, the objectives should ensure the quality of the framework. Semi-structured interviews help 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.

Phase 3: Design and Development. (Chapter 4)

[SQ3]: How can the objectives be transformed into design-requirements?

Within this phase, the objectives [SQ2] will be transformed through specification, as described by van de Poel (2013). Furthermore, the contextual needs for the framework will be transformed into design-requirements. Lastly, semi-structured interviews will improve 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 deliverable of this question will consist of a list of design-requirements for the framework.

[SQ4]: How can the design-requirements be transformed into a pragmatic prospective assessment framework?

As next step within this phase, SQ4 will be answered. This will be done by creatively converting the design-requirements into an assessment framework. This development phase results in the assessment framework that can be used and has the characteristics that the introduction of chapter 2.5.2 outlines.

Phase 4: Demonstration. (Chapter 5)

[SQ5]: Is the framework pragmatic to use for assessment of explicability with a specific case?

This phase will focus on using the framework to assess explanations, produced by different techniques (explanation techniques), within the case of interest (credit underwriting) and demonstrate the pragmatic level of it. The choice for which techniques to include will be based on the ability to show that the framework achieved the objectives. The input for this is the assessment framework and the result of this phase is a conclusion whether the case that is being assessed is explicable enough, with the specific context that is applied, to be deployed. The used framework additionally provides the justification for the conclusion, since it can be traced back if all criteria are met, or which criterium is not met. This demonstration shows how the assessment framework should be implemented with an example case, and how it can be used for decision making concerning the sufficiency of explicability.

Phase 5: Evaluation. (Chapter 6)

[SQ6]: Does the demonstration show that the framework accomplished the objectives?

This phase evaluates the outcomes of the demonstration phase and answers SQ5, i.e. are all formulated objectives achieved? This will conclude in the limitations and knowledge of the achievements of the framework. If the objectives are not sufficiently met, the decision-makers can decide to iterate back to phase 3 and redesign the framework. If they are met, one can continue towards the last phase and communicate the results of the design, demonstration and evaluation phases.

Phase 6: Communication. (Chapter 7)

The last phase consists of the discussion on the research phases, results and process. Further, it includes guidelines for using the framework (appendix 3) and the generalization possibilities will be discussed as well. Chapter 7 is the last chapter, which forms the conclusions, reflects on the research & these conclusions, and formulates recommendations for future research. This includes the answer to the main research question, and it outlines the general knowledge that is obtained through this research. In

addition, the relevance of the designed framework in solving the problem (chapter 1) is the main topic of this thesis, and it simultaneously is the way to communicate the results of the research.

2.5.3 Research flow diagram

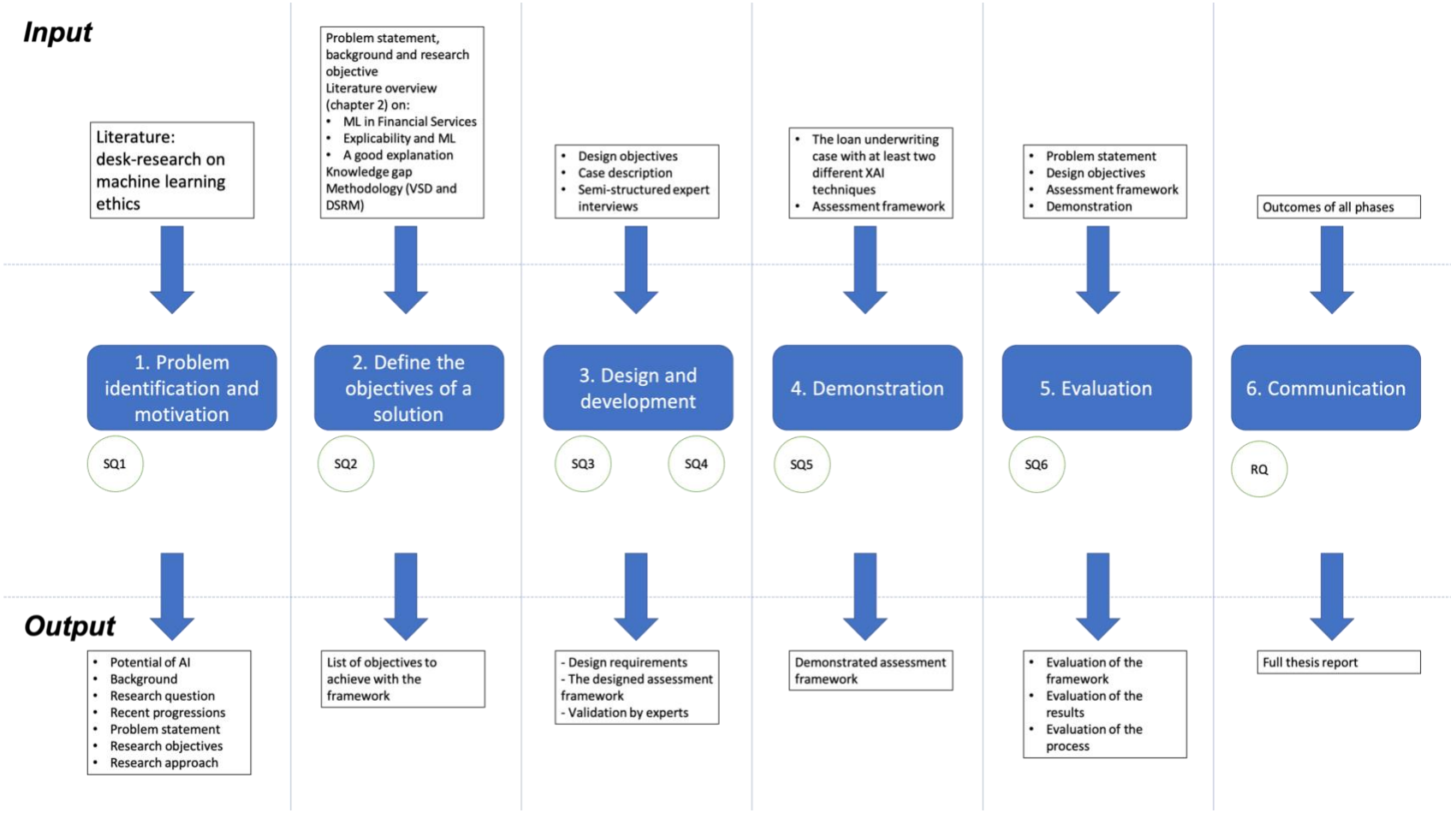


Figure 9: Research flow diagram

2.6 Summary of chapter 2

Summarizing, this thesis will focus on the use of machine learning systems for credit underwriting with personal consumer loans in the European Union. Explicability is the value of interest, as a means to accountability, that subsequently is a means to fairness and trust, visualized in figure 11.

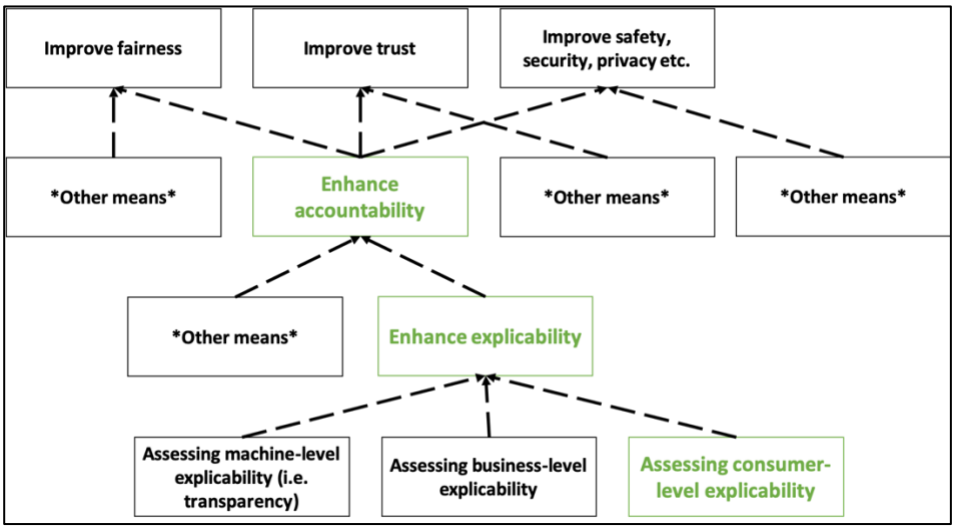


Figure 10: Conceptual relation - Means-End diagram

An *explicable* machine learning 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. Since *transparency* concerns the possibility to inspect the workings of algorithms and the data used, this can be seen as a means to explicability. *Accountability* is a social relationship and is defined as 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.

In addition, there is the regulator perspective. The Consumer Credit Directive (CCD) and the General Data Protection Regulation (GDPR) are regulatory drivers for explicability of machine learning systems. Further, there exists an ethical and societal need for the enhancement of explicability.

To accomplish this, the CRISP-DM lifecycle model will be expanded with an extra dimension of evaluation of the machine learning system: the explicability perspective. A framework will be designed to assess the explicability of a machine learning system on consumer-level. This framework needs to be pragmatic for decision-makers in the development lifecycle (the model developers, data scientists), and to be used for different techniques that formulate an explanation (model-agnostic).

The framework shall be able to prospectively assess the goodness of an explanation as a product. Within this thesis, local ex-post explanations with a justificatory value for a layperson are in scope. To do a good assessment, the framework should include the four identified principles for good explanations: soundness, completeness, comprehensibility and conciseness. It should also be adapted to the context characteristics of the case.

The scientific goal of this thesis is to fill the identified knowledge gap. There lacks research on how to prospectively assess explicability on a pragmatic level, on the multi-disciplinary approach of the goodness of explanation and thus the contextual case characteristics that need to be incorporated in the framework, and literature lacks a robust framework that is incorporated in a development lifecycle so that it is multi-organizational as well.

Lastly, this thesis makes use of the Design Science Research Methodology complemented by the Value Sensitive Design Approach as a structure and research method in order to complete this goal.

3. Chapter 3. Framework objectives

Within chapter 2, the literature overview provides the concepts and relations relevant to the explicability of machine learning with credit underwriting. Further, the scope has been set, the first criteria for a good explanation are outlined, the knowledge gap has been determined and the methodology is described. We proceed in chapter 3 with the second phase of DSRM (objectives for the solution) where we will formulate the objectives for the framework to be designed. This formulation will be based on the problem statement (chapter 1.3), the research objective (chapter 1.5) and the literature overview (chapter 2). It will answer the second sub-question [SQ2]: *What are the objectives for the assessment framework?* The definition of the objectives for the assessment framework has three final goals:

- Creates a starting point for the formulation of design requirements for the framework
- It should be able to ultimately evaluate the designed framework (or multiple frameworks if the aim is to compare multiple ones) on these objectives (phase 5 of DSRM)

The first paragraph will describe some first assumptions that are needed to be elaborated upon, before the objectives can be formed, in order to set the scope. The second paragraph will outline the objectives that are derived and describes these objectives. This chapter finalizes with a short summary of the chapter.

3.1 Assumptions for the formulation of the objectives

The first assumption is that this framework concerns the assessment of textual explanations: so not visual or verbal explanations. The reason for this is that visual explanations open up a whole new field of cognitive science, vision science, and visualization that this thesis does not aim to cover. In addition, verbal explanations do increase the complexity, with aspects like the clarity of speech and hand gestures, and are therefore more related to the social process of explanation. Moreover, in order to verbally communicate an explanation to someone, one can argue that it should be able to write that explanation down before you are able to do that and so we end up back at the textual explanation.

The second assumption is derived from the explanation type scope: it solely concerns the assessment of local ex-post explanations as a product. As we focus on prospective assessment, this implies that the to-be-assessed local ex-post explanations are hypothetical at that time (see figure 12); the explanation is on hypothetical decisions after deployment (of version 1) of the machine learning system, where deployment is at time T = N, assessment at T = N – 1 and the decision with corresponding

explanation at T = N + 1. This scope influences the refinement of certain objectives though (objective D, E, F, H and I), and chapter 4.1 (design requirements) will discuss this in more detail.

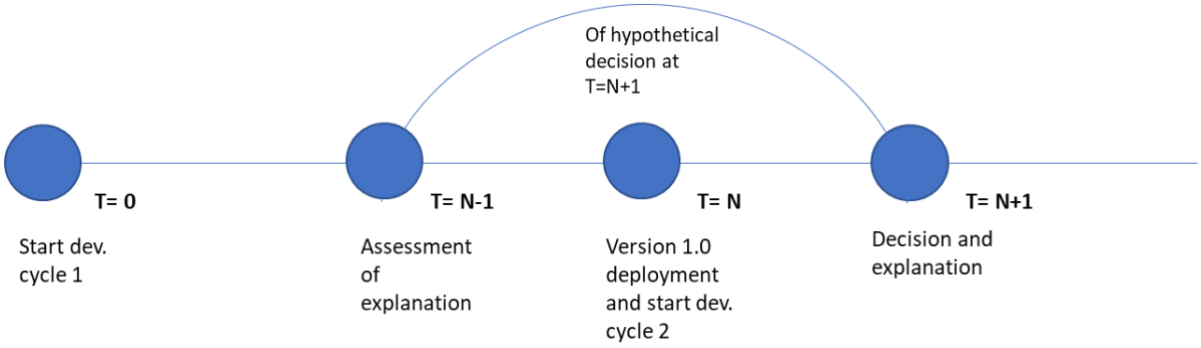


Figure 11: Timeline - a hypothetical decision

The third assumption is that the decision-makers within the development lifecycle try to maximize the prediction-accuracy of their machine learning system while sustaining a sufficient level of explicability for the use case. The main advantage of using a more complex machine learning system is the improvement in prediction-accuracy or performance, however, for the case of interest, the machine learning system needs to sustain a certain level of explicability. So, the assumption is that the decision-makers will choose a system that has better performance with a sufficient level of explicability over a system that has lower performance and also a sufficient level of explicability.

3.2 Design objectives

The design objectives should safeguard the quality of the framework, the scope and the focus on the explicability-performance tension. The objectives are in addition the top-level goals that the design requirements are derived from (chapter 4).

We approach the design of the framework from a *VSD* perspective, as opposed to a *Design Optimization* approach. VSD is a multi-objective approach, instead of a mono-objective approach, and this approach fits our case the best; VSD is focused on bringing in values and objectives from different stakeholders and show multiple design options, such that finally a sufficient design can be chosen (Friedman et al., 2013). Firstly, machine learning systems in large organizations like banks are designed and used by multiple people and have multiple stakeholders. Secondly, the ethics of one singular person is not addressed, but the ethics from a society. Thus, how could it be that a machine learning system, developed by, used by and impacting various different people and stakeholders with diverse values has just one objective? It is misleading to say that the system only has one objective and to design the system

accordingly. It is more practical to do this (in the sense that it is easier to commit to one objective), however, it is less valid in relation to real-world situations. Therefore, multiple objectives are formulated, which could result in a Pareto optimal solution with the design of the framework. If no objective-based argument can be stated why a certain design is superior over another, a choice has to be made. In this situation the final decision on which framework design to adopt will be made on subjective preference by the designer of the framework in this thesis, since multiple solutions are considered equally good.

The complexities that are derived from the problem statement, in combination with the research objective and the literature overview, are transformed into design objectives (table 4). A reference where the complexity is discussed is added in brackets. The following design objectives are the ones that the to-be-designed framework needs to achieve:

Table 4: Design objectives for the to-be-designed framework

Complexities to solve	**Design objectives**
It is unclear for decision-makers in the development lifecycle how to structurally evaluate explicability of machine learning systems. Further, it is not clear how to decide on whether the system is sufficiently explicable; the framework needs to be pragmatic. (chapter 1.3)	A. Provides guidance for the decision-makers on using the framework
	B. Helps with decision-making on whether a certain task or decision-making functionality can be delegated to the machine learning system
Principles should be made operational and data scientists in the development lifecycle need to be able to include *design for explicability,* but it is currently interpretable in how to do this. (chapter 1.3, 1.6 and 2.2.3.1)	C. Is able to prospectively assess explanations
There is currently no framework that is specified (or adjusted) to the context of the case (accountability as a value, CCD, GDPR, and ethical need), whereas this is necessary. (chapter 2.2.1.1, 2.2.1.3 and 2.2.2.1)	D. Is able to assess justificatory explanations
	E. Is able to assess consumer-level explanations towards a layperson
There is literature available on qualitative criteria concerning '*good'* explanations, and techniques that can provide explanations and improve the explicability of machine learning systems. However, there is currently no framework that incorporates the structural assessment of	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

explanations on these qualitative criteria that an explanation should possess. (chapter 2.3)	

3.2.1 Objective A: Provides guidance for the users of the framework

In order to move from the theoretical field of values towards the operationalization of these values into the system, the designers of the machine learning system must know how they should implement these values. It should be clear for the users of the framework how they can incorporate tasks in the development lifecycle. This will entail new tasks, where it should be known which tasks, how, in what sequence and at what moments these tasks need to be performed. This needs to be tuned with the CRISP-DM lifecycle.

If the proposed users cannot use the framework with the user's guidelines, this objective is not reached. For the pragmatic perspective this is a very important objective.

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

The main question that decision-makers within the development lifecycle want to answer is: "Is the current designed machine learning system explicable enough to be deployed within this specific context?" First, this is dependent on the explicability scope that the framework is built for, so a final decision at this question can only be made if the full explicability scope is assessed. Nonetheless, the assessment framework can be used to be able to assess one part *(local ex-post justificatory explanations at consumer-level towards a layperson)* of the full scope, and whether the machine learning system is explicable enough for this a first step is made. In order to help this decision-making, the framework should be able to be used to combine the results of the assessment to form an answer to the question. If the answer is negative, the assessment framework should be able to show the decision-makers what explicability issues need to be solved in order to improve the explicability to a sufficient level. In addition, the decision-makers need to know to which phase they have to iterate to try to accomplish this.

If those aspects aren't included in the conclusion in the framework, it is not sufficiently reached. For the pragmatic perspective this is an important goal to accomplish.

3.2.3 Objective C: Is able to prospectively assess explanations

The fact that we focus on *explicability by design* (Value Sensitive Design), implies that we take a prospective viewpoint. Moreover, we look at assessment as a way to evaluate to what level the machine learning

system is explicable (to give designers guidance for improving the explicability with the design), which supports this view as well.

Prospective assessment can be defined as an assessment that will take place before the system is deployed. Since this prospective assessment is meant to help designers to decide if the right explanation producing technique has been chosen (i.e. is the produced explanation good enough), the framework needs to be *technique-agnostic*, so that different techniques can be compared on the product that they produce.

If no hypothetical decision can be assessed with the framework, this objective is insufficiently reached. From the VSD perspective this is foremost an important objective to suffice.

3.2.4 Objective D: Is able to assess justificatory explanations

Concerning the aim to improve accountability by improving explicability, and to take on the goal to produce explanations that are the most valuable for the loan applicants, the focus is on *justificatory explanations* opposed to *teaching explanations.* This concerns an explanation that answers the justification question: *"Why should we believe that this prediction is correct?"* (Biran & McKeown, 2014). '*Correct'* can be interpreted more broadly than just correct in the mathematical sense: e.g. consumers could as well be interested to know that just ethically and legally accepted variables are used for the prediction. So, the explanation needs to include evidence for the correctness of the decision, and the framework needs to be able to assess the explanation on this.

If the explanation cannot be assessed on the aspects that form the evidence for the decision, this objective is not reached. Especially for the accountability value this is a very important objective.

3.2.5 Objective E: Is able to assess consumer-level explanations towards a layperson

Additionally, to this justification aspect, there is the human-understandability of an explanation. The framework will be used to assess *consumer-level explanations towards a layperson*. Therefore, the explanation needs to be assessed on what level of knowledge the explainee requires to have in order to understand the explanation.

This objective is partly dependent on the social process of explicability as well and should be incorporated in the future design of an assessment framework of this process, but this thesis does not cover this social process.

It is an important aspect to incorporate since the satisfaction and value of these types of explanations depend on this. Further, it creates a tool for to the needs as defined by GDPR and CCD

3.2.6 Objective F: Is able to assess the completeness of explanations

Moving towards the assessment criteria for a good explanation (Kulesza et al., 2015, 2013), the explanations need to be assessed on the required level of completeness that it needs to have. To do this, a first selection needs to be made of intelligibility types (Lim & Dey, 2009) that are relevant for the explanation of interest. Next, the framework must be able to assess whether the explanation includes the intelligibility types that are considered relevant to be included.

If the framework does not include a tool to do the assessment, this objective is not reached. Completeness and conciseness have a tension, so the importance of these criteria relative to each other depends on the explainee needs (this should be empirically investigated in future research).

3.2.7 Objective G: Is able to assess the soundness of explanations

The soundness of an explanation concerns the following question: 'Is the explanation a correct representation of how the system made the decision?' So, within the consumer-level explanation, it does not look after the validity of the model, but after the validity of the explanation with respect to what is known about the model. The truthfulness of an explanation with respect to the internal workings of the system will be assessed here.

In order to fully ensure the soundness of the complete machine-learning system, it is very important that the soundness of the machine-level explanations and business-level explanations are assessed and truthful, regarding this aspect. Since this assessment area is out-of-scope for this thesis, this is a very important area that has to be investigated in future research.

The framework needs a tool to assess this criterion in order to reach this objective. The soundness is an essential aspect for all explanations, since it is evident that the explanations need to be truthful.

3.2.8 Objective H: Is able to assess the comprehensibility of explanations

The comprehensibility of an explanation concerns the human-understandability of explanations like in objective E, however, it slightly differs. Objective E concerns the definition and acceptance of the concepts that can be used in the explanation, in order to keep a sufficient level of understandability. This objective takes these concepts and concerns the linguistic aspects of formulating an understandable explanation with these concepts; e.g. contextual details around singular facts, narrative format, the language and the logical relation between decision, important features, roles, effects, and important values. The framework must include these aspects for assessment to reach this objective. Concerning the GDPR, the CCD and the layperson perspective this is especially an important objective.

3.2.9 Objective I: Is able to assess the conciseness of explanations

The last objective concerns the conciseness of an explanation. An explanation can be sound, complete and comprehensible, but when an explanation is very lengthy, the explainee will be distracted, less interested and this could lead to less satisfaction and trust in the explanation. Therefore, an explanation needs to be assessed on the number of textual lines, words and concepts there are in the explanation. In addition, it is helpful if an explanation has a modular structure in such a way that it can easily be extended when consumers ask for more reasoning and justification, without losing the structure of the explanation. Therefore, more and less demanding consumers can be served. As well as objective E, this objective relates to the social process: an explainee has to ask for more explanation, in order for the explainer to know that the explanation needs to be extended.

This objective has a tension with the completeness of an explanation: in order to completely describe certain features of a system, the explanation sometimes has to be longer. We argue that a good balance needs to be aimed for. It is important that the framework includes handles to assess this aspect in order to reach this objective, but it is dependent on the explainee to what level the explanation needs to be concise.

3.3 Summary of chapter 3

The goal of this chapter was to answer sub-question 2: *What are the objectives for the assessment framework?* To do this, three important assumptions are outlined first:

- The assessment will solely focus on textual explanations
- 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
- 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 statement, the research objective and the literature overview. The design objectives are transformed from the complexity in such a way that the accomplishment of the objective(s) solves the complexity. 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.

4. Chapter 4. Explicability Assessment Framework Design

The last chapter formulated the objectives that the framework needs to achieve. We proceed with the design and development phase (phase 3 of DSRM), where design requirements will be derived from the objectives. These design requirements will be used for the development of the framework. Within this chapter, two sub-questions will be answered: "*How can the objectives be transformed into design-requirements?"* [SQ3] and "*How can the design-requirements be transformed into a pragmatic prospective assessment framework?"* [SQ4]. This chapter will finalize with a short summary.

4.1 Design Requirements

To transform the objectives into design requirements ([SQ3] *how can the objectives be transformed into design-requirements?*) we refer to Van de Poel (2013), who describes the process of specifying values into norms into design requirements. The design requirements should serve a '*for the sake of*' relation with the upper-norm, or objective; this relationship will be leading the formulation of the design requirements.

A requirement can be classified as a member one of the following three categories: *functional requirement, non-functional requirement* and a *constraint* (Faulconbridge & Ryan, 2014, p.47). A functional requirement (FR) is "*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) is "*a quality, property or attribute that the system must possess*". A constraint (C) is *"a restriction or bound under which the system should operate, or the way in which the system is to be developed"*. The system as mentioned in the definitions can be seen as the artifact to-be-designed: the framework. A system is defined as *"a combination of interacting elements organized to achieve one or more stated purposes"* (Faulconbridge & Ryan, 2014, p.3). The interacting elements are the questions within the framework that form the assessment by the user of the framework, and the purpose is to form ultimately a conclusion on the level of explicability of the machine learning system, by using this framework. The derived design requirements will be classified according to these three categories.

Design requirements have a level of specificity. When approaching this from the axiomatic design perspective (Suh, 1998), we can observe two things. First, our defined objectives are from the customer domain; the customer needs. The 'customer' here is the decision-maker who will execute the explicability assessment. Second, when we take a step further to the design requirements, we see that the objectives are mapped to these design requirements. Moreover, the design requirements reflect "*how we propose to satisfy the requirements specified in the left domain*" (Suh, 2018 p. 204) and the formulated design

requirements are in the functional domain of the 'design world'. This has implications for the design requirements, and this will be discussed in chapter 4.1.10. Furthermore, in this chapter a design requirements interrelation matrix is created. This matrix shows an overview of whether the design-requirements interrelate. In chapter 4.1.1 to 4.1.9 the rationale for the design requirements is given and they are discussed with regards to the interrelation with other design requirements.

To create clarity, in the next chapters the interrelations that are already discussed in a section of a des. req. won't be repeatedly discussed after that. As an example: the des. req. interrelation with the user's guidelines will be discussed in 4.1.1. but not in the sections after that.

4.1.1 Design requirements to objective A

Objective A: Provides guidance for the decision-makers on using the framework

To enhance the pragmatism of the framework, "how-to-use"-guidelines need to be attached to the framework as a sub-part of the framework. These instructions should inform the decision-makers on how to use the framework and guide them. The following design requirements are formulated to achieve this:

- **[FR.A1]** shall show which new tasks need to be performed in the development lifecycle

Rationale and interrelation: The user of the framework has to know what tasks to perform. The tasks to perform interrelate with all the aspects of the framework, since this covers what has to be done. If the framework changes, the tasks to perform changes.

- **[FR.A2]** shall show how the new tasks need to be performed in the development lifecycle

Rationale and interrelation: The user of the framework has to know as well how a certain task needs to be performed. If the framework changes, the how-part of the guidelines changes as well, so this interrelates with all the other design requirements as well.

- **[FR.A3]** shall show in what sequence the tasks need to be performed in the development lifecycle

Rationale and interrelation: The tasks have a certain sequence to follow in order to create the most value with the framework, and users need guidance for this. The sequence of the tasks to perform is dependent on the des. req. that relate to the adaptation of the framework to the context (FR.D1, FR.D2, FR.E1, FR.E2, FR.F1 and FR.F2). This influences the user's guidelines as well. This needs to happen before the actual assessment by using the framework.

- **[FR.A4]** shall show at what times in the development lifecycle the tasks need to be performed

Rationale and interrelation: The framework needs to be aligned with CRISP-DM, so that users are guided with when to perform the assessment. This solely interrelates with the des. req. that concern the process of assessing opposed to the assessment itself (FR.B1, FR.B2 and C.C1).

4.1.2 Design requirements to objective B

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

As mentioned, the final purpose of using the framework is to conclude in a decision if the machine learning system is explicable enough (within the context of the case and the explanation of interest) to be able to take on the decision-making on the acceptance or denial of a loan request. Further, it needs to be clear what aspects cause issues of inexplicability if that is the case so that these issues can be tried to be solved. Thus, the framework:

- **[FR.B1]** shall result in a conclusion if the explanation of interest is good enough

Rationale and interrelation: A conclusion improves the pragmatic relevance of the framework for users. This des. req. interrelates with the all the des. req. *except* the ones that concern the adaptation of the framework (FR.D1, FR.E1 and FR.F1), because a conclusion is based on the intermediate steps of the framework.

- **[FR.B2]** shall give an overview of the specific explicability issue(s) in the system if it is not explicable enough

Rationale and interrelation: The decision-makers that use the framework need next steps to move forward or iterate back with, when something is wrong. This FR interrelates to the design requirements that concern the choices and discovery of explicability issues in an explanation (FR.D1, FR.D2, FR.E1, FR.E2, FR.F1, FR.F2, FR.G1, FR.H1, FR.H2, FR.I1, FR.I2 and FR.I3)

4.1.3 Design requirements to objective C

Objective C: Is able to prospectively assess explanations

To align with the VSD approach, the assessment framework is focused on the pre-deployment design phase, before actual ethical problems with consumers could potentially occur. In addition, the users of the framework should be able to assess the explanation products that are formed by different techniques in order to evaluate whether another technique formulates a better explanation.

- **[C.C1]** shall be usable for explanation assessment before the machine learning system is deployed

Rationale and interrelation: This requirement is important to be able to ensure that an assessment has taken place before problems in the real world have occurred. This constraint does not interrelate with further design requirements (except the ones mentioned before), due to the fact that an assessment after deployment should not result in different FR.

- **[C.C2]** shall be technique-agnostic, thus usable to assess different formats of explanations

Rationale and interrelation: Since there are multiple methods to create explanations, and different explanation formats, the framework needs to be able to deal with this. When this is possible, the decision-maker can prospectively choose a technique that creates the right explanation and directly assess this explanation. This des. req. interrelates with the FR that concerns explanation specific aspects; i.e. not the length of an explanation (FR.I1, FR.I2).

4.1.4 Design requirements to objective D

Objective D: Is able to assess justificatory explanations

To be able to assess the justificatory aspect of the explanation (to align the framework with the goal of improving accountability), we make use of the concept of *evidence roles in justificatory explanations* as described by Biran & McKeown (2014). The framework should be able to be used to check which evidence roles fit the needed explanation and should further be able to check which evidence roles are included in the explanation that is being assessed.

- **[FR.D1]** shall be able to check which 'evidence roles' are useful to be included in the explanation

Rationale and interrelation: This sets a threshold for the justificatory value that the explanation needs. This FR interrelates with the content focused des. req. since the choices here could influence the content (FR.D2, FR.E2, FR.F1, FR.G1, FR.I1, FR.I2 and FR.I3).

- **[FR.D2]** shall be able to check which 'evidence roles' are present in the explanation

Rationale and interrelation: A check needs to take place to find out if the required evidence roles are included in the explanation, and to check if these could imply any problems with the ML system. The evidence roles suffice in the justificatory value. This design requirement interrelates with the FR concerning the knowledge level of the explainee (FR.E1) and the soundness (FR.G1), since the evidence roles that are included in the explanation influences the required knowledge level and the ability to check the soundness.

4.1.5 Design requirements to objective E

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

The framework needs to be adjusted to the type of explanation as well and be able to evaluate the human-understandability of the explanation. To provide a tool for this, the framework needs to include an evaluation of the required knowledge level for the explainee to understand the explanation. Additionally, the framework should be able to answer whether the explanation is understandable for a layperson.

- **[FR.E1]** shall be able to specify the knowledge level of expertise that the explainee needs to have in order to understand the explanation

Rationale and interrelation: To have a threshold for the understandability of the explanation, it first needs to be clear what the knowledge level of the explainee needs to be for the explanation. This FR interrelates with the design requirements that concern the content of the explanation (FR.E2, FR.F1, FR.G1, FR.H1 and FR.H2), because this influences the required knowledge level.

- **[FR.E2]** shall be able to conclude whether the explanation is understandable enough for a layperson

Rationale and interrelation: With regards to the knowledge level of expertise of the explainee, a check needs to take place to find out if the explanation is sufficiently understandable. This des. req. does not further interrelate with further FR.

4.1.6 Design requirements to objective F

Objective F: Is able to assess the completeness of explanations

Lim & Dey (2009) provide a comprehensive list of different intelligibility types that an explanation could include. However, the purpose, explainee, and context of the explanation eventually define which types are required to be included. This is the first step that needs to be included: check which types need to be included in order to have enough completeness of the explanation. The next logical step is to assess which intelligibility types are included in the explanation.

- **[FR.F1]** shall be usable to check which intelligibility types the explanation requires to fulfill a sufficient level of completeness

Rationale and interrelation: This design requirement sets a threshold that the explanation requires to have a sufficient completeness. This FR interrelates with all the further des. req. (FR.F2, FR.G1, FR.I1, FR.I2 and FR.I3) except for the ones that concern the comprehensibility (an explanation can still be in narrative format and understandable). The intelligibility types that need to be included define the content of the required explanation. This influences the length and modularity of an explanation as well.

- **[FR.F2]** shall be usable to assess the extent to which the relevant aspects are included in the explanation

Rationale and interrelation: A check needs to take place which of the required intelligibility types are included, to assess completeness. This des. req. is not further interrelated with other FR.

4.1.7 Design requirements to objective G

Objective G: Is able to assess the soundness of explanations

The soundness concerns the truthfulness of the explanation in relation to the underlying system (Kulesza et al., 2015, 2013). "*The more the explanation reflects the underlying model, the more sound the explanation is"* (Kulesza et al., 2015, p.127). This can be evaluated by assessing the different components of an explanation on the validity of these components with the underlying system; is the explanation a correct representation of how the model made the decision?

- **[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

Rationale and interrelation: An explanation needs to be based on the truth, and this assessment helps the check for soundness of the explanation. This des. req is interrelated with all the further FR (FR.H1, FR.H2, FR.I1, FR.I2 and FR.I3). It can be possible that the more truthful an explanation is, the harder it is to have a narrative format and include solely easily understandable human-interpretable language. The truthfulness of an explanation can influence the length and modularity of an explanation as well.

4.1.8 Design requirements to objective H

Objective H: Is able to assess the comprehensibility of explanations

In order to enhance the comprehensibility of explanations, the narrative aspects of an explanation need to be assessed. This narrative viewpoint is important to improve the intelligibility of an explanation (Biran & McKeown, 2014; Velleman, 2003). In addition to the narrative aspect of explanations, the linguistic aspect of textual explanations is considered important as well, especially in the context of consumer-level and a layperson as explainee; an explanation needs to be understandable. Thus, the framework:

- **[FR.H1]** shall be usable for the assessment of the narrative aspect of an explanation

Rationale and interrelation: A narrative format increases the intelligibility of an explanation and thus the comprehensibility. This FR is interrelated with the other further des. req. (FR.H2, FR.I1, FR.I2 and FR.I3). How easy it is to have a narrative format is dependent on the language used, and it influences

the length of the explanation, how many concepts are included and especially how easy it is to have modularity.

- **[FR.H2]** shall be usable for the assessment of the human-interpretable linguistic aspects of the explanation

Rationale and interrelation: Since the explainees are human laypersons, the choice of words is an important aspect and these need to be understandable for them. This des. req. is unrelated to further FR, since whatever length an explanation is it, or how modular it is, it can still include solely easily understandable language.

4.1.9 Design requirements to objective I

Objective I: Is able to assess the conciseness of explanations

Lastly, Kulesza et al. (2015) mention that a more concise explanation could enhance the understandability of the explanation. Therefore, the length of an explanation should be assessed. This characteristic must be assessed relative to the completeness of an explanation since there is a tension here. The question to consider is: can we have the same level of completeness with better conciseness?

Further, in order to have options to give the explainee a more comprehensive explanation if requested, it would be valuable if the explanation is easily extendable. Modularity (the ability to add or separate a component of explanation to the former explanation) of an explanation enhances the ability to extend an explanation without changing the structure of the explanation.

- **[FR.I1]** shall be able to assess the explanation length

Rationale and interrelation: The shorter the explanation length, the more concise the explanation is. This FR is related to the number of concepts des. req. (FR.I2): the more concepts the longer the explanation.

- **[FR.I2]** shall be able to assess the number of concepts included in the explanation

Rationale and interrelation: The less concepts (the words that carry real value for the meaning of the explanation) there are in the explanation the more concise the explanation is. Note that a sufficient number of concepts need to be included to have sufficient meaningful value for the explainee. This des. req. is unrelated to the modularity FR, since an explanation can have a lot of concepts and still be very modular, or not at all.

- **[FR.I3]** shall be able to assess the modularity of the explanation structure.

Rationale and interrelation: The more modular the explanation is, the easier it is to make an explanation more concise without losing structure nor making the explanation more complex. Furthermore, if an explainee asks for more information, this could more easily be added. All the interrelations of this FR are already discussed in earlier sections.

4.1.10 Design requirements matrix

As can be seen in chapter 4.1.1 to 4.1.9, there are quite some interrelations between the design requirements. The interrelations are documented in table 5 as well where every 'X' is an interrelation.

An optimal design within the axiomatic design approach commits to two axioms (Suh, 1998): *the interdependence axiom* and *the information axiom.* The first means that it is desired to have FR that can be satisfied without affecting other FR. The second means that it is desired to have a design that *minimizes the information content* or maximizing the probability to satisfy a certain set of FR.

It is evident that with the current design requirement interrelations this is hard to accomplish, however I argue that this is not a problem for two reasons. First, as chapter 3.2 described, we approach the design of the framework not from a *design optimization approach*. It needs to be a guiding framework that is sufficiently good concerning the important objectives. Second, the axiomatic design as described by Suh (1998) focuses on a physical system opposed to our conceptual system of a framework. Furthermore, it focuses more on mono-disciplinary systems (e.g. manufacturing, software) as opposed to the multi-disciplinary system of our framework.

Nonetheless, the matrix notation from the axiomatic design approach is an interesting way to show the interrelations and the tensions between the design-requirements, and the mapping of the objectives and requirements on the different layers as defined by Sub (1998) gives a new perspective to these. Future research could as well be conducted to investigate to what extent this framework could have been developed differently with regards to using the axiomatic design approach.

Table 5: Design requirements interrelation matrix

	FR.A1	FR.A2	FR.A3	FR.A4	FR.B1	FR.B2	C.C1	C.C2	FR.D1	FR.D2	FR.E1	FR.E2	FR.F1	FR.F2	FR.G1	FR.H1	FR.H2	FR.I1	FR.I2	FR.I3
FR.A1		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
FR.A2	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
FR.A3	X	X							X	X	X	X	X	X						
FR.A4	X	X			X	X	X													
FR.B1	X	X		X		X	X	X		X		X		X	X	X	X	X	X	X
FR.B2	X	X		X	X				X	X	X	X	X	X	X	X	X	X	X	X
C.C1	X	X		X	X															
C.C2	X	X			X				X	X	X	X	X	X	X	X	X			X
FR.D1	X	X	X			X		X		X		X	X		X			X	X	X
FR.D2	X	X	X		X	X		X	X		X				X					
FR.E1	X	X	X			X		X		X		X	X		X	X	X			
FR.E2	X	X	X		X	X		X	X		X									
FR.F1	X	X	X			X		X	X		X			X	X			X	X	X
FR.F2	X	X	X		X	X		X					X							
FR.G1	X	X			X	X		X	X	X	X		X			X	X	X	X	X
FR.H1	X	X			X	X		X			X				X		X	X	X	X
FR.H2	X	X			X	X		X			X				X	X				
FR.I1	X	X			X	X			X				X		X	X				X
FR.I2	X	X			X	X			X				X		X	X				
FR.I3	X	X			X	X		X	X				X		X	X		X		

4.2 Framework development

The development of the assessment framework will be performed in two sequential steps. First, potential means for the formulated functional requirements are drafted, after which one (or multiple) will be chosen in such a way that all the requirements are met. The second step is to synthesize and implement the means into a framework that can be used in practice, and answer sub-question 4 [SQ4}: *How can the design-requirements be transformed into a pragmatic prospective assessment framework?*

4.2.1 Means for meeting the design requirements

Continuing with the formulated design requirements, we can observe that most requirements are functional requirements that thus focus on the fulfillment of "*a service or function that the system should provide, a thing it should do, or some action it should take"*. The combinations of functional requirements on the vertical axis of a matrix with possible means to meet these requirements on the horizontal axis of a matrix, result in a *morphological chart* (Dym, Little, & Orwin, 2014). It effectively represents the design space that we are working in and we select at least one means from every row, in order to choose and develop a possible design (table 6). The means are the tools that could be included in the framework to accomplish meeting the functional requirements.

Table 6: Function means diagram (Morphological chart) for design generation

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	
[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?	

4.2.2 Means synthesis: Explicability Assessment Framework (EAF) creation

The means of table 6 need to be synthesized into a usable assessment framework. Before we move to this step, figure 13 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.

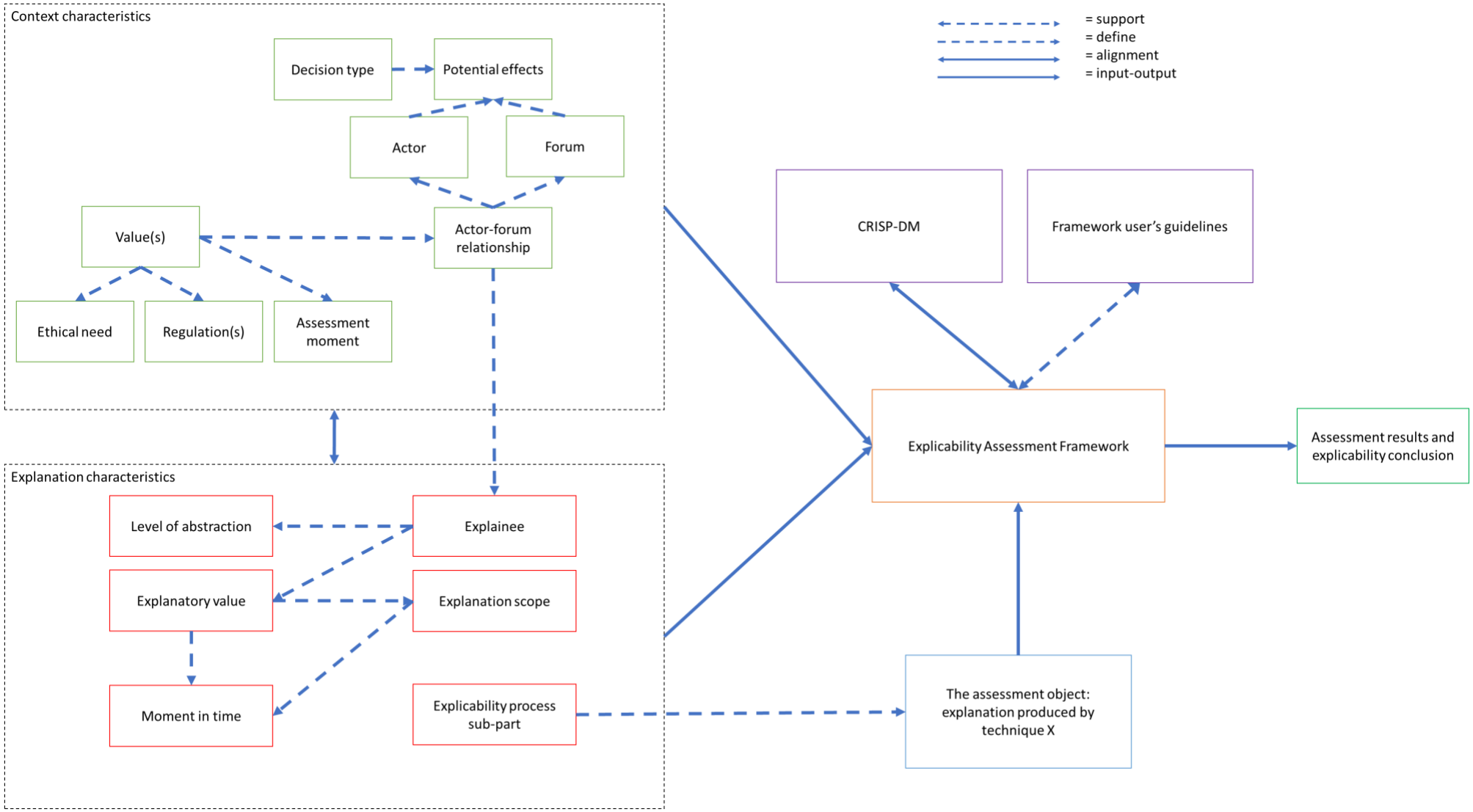


Figure 12: Assessment framework relations

Table 7 shows the Explicability Assessment Framework (EAF) that is derived from combining the means from table 6 and the relations as drafted in figure 13. Furthermore, input from industry experts, gathered by conducting semi-structured interviews (appendix 2.2), is used for steering and validation of the design of the framework. Chapter 4.3 will summarize the meaningful aspects from these interviews and explain what it meant for the design process.

User's guidelines are designed for the EAF with the goal to create an optimal understanding of how to use the framework, and to enhance the pragmatic aspect of the framework. These guidelines are included in appendix 2.3 and will guide the user with the use of the framework to assess a use case. This will be shown in the demonstration (Chapter 5).

The proposed user of the framework and the user's guidelines is the decision-maker within the development lifecycle. Designers of the machine learning system are considered decision-makers here, as they are required to make important decisions or choices along with the development of the system that impacts the performance and explicability of the system. However, more senior designers are assumed to have more 'decision-power' concerning the choices of within the design of the machine learning system; they have the final vote for or against a model, technique or design. Therefore, the assessment framework can be seen as the most valuable for the senior designers within the development cycle, or the data-scientist managers.

Table 7: Explicability Assessment Framework (AEF)

1. <u>Context characteristics</u>	
Decision type	*insert description of the type of decision that the machine learning system makes in the case of interest*
Value(s)	*insert the value of interest for the design of the machine learning system in the use case*
Actor-forum relationship	*insert a description of the type of relationship the actor and the forum have*
Forum	*insert the forum of the decision*
Actor	*insert the actor of the decision*
Potential effects	*insert a description of the potential effects for the forum and actor of the decision*
Ethical need	*insert a description of the ethical need that drives incorporating the value in the design of the system*
Regulation(s)	*insert a description of the regulation(s) that drives incorporating the value in the design of the system*
2. <u>Explanation characteristics</u>	

Explainee	Layperson/business/data-scientist/auditor	
Level of abstraction	Consumer-level/business-level/machine-level	
Explanatory value	Justification/teaching	
Explanation scope	Local/global	
Moment in time	Ex-post/ex-ante	
Explicability process sub-part	product/cognitive process/social process	
3. <u>Framework adjustments</u>		
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	
4. <u>*Assessment Object* Assessment</u>		
insert the full explanation to be assessed		

Questions:	***Answers:***
Justificatory explanation	
1.X For evidence role X, does the explanation contain this evidence role?	
Explanation towards a layperson	
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?	
Completeness	
3.X For intelligibility type X, does the explanation contain this intelligibility type?	
Soundness	
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?	
Comprehensibility	
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?	
Conciseness	
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. <u>Concluding section</u>	
*insert conclusion … *	

4.3 Expert validation objectives & design requirements

Appendix 2 contains three conducted semi-structured interviews. The goal of these was to create more validity with regards to the objectives and design requirements of the framework.

The first interview shows that the formulated objectives (at that moment) are relevant and useful for the practice and that these could be used for a framework. It as well provided relevant feedback for the adjustments later of the objectives.

The second interview showed complications with the understanding of the goal and the context of the framework. Since this is a qualitative framework and the respondents were quantitative experts, the

hypothesis is that this caused a mismatch of interest and knowledge within this field. This led to interesting insights on the industry applicability of Value Sensitive Design in a quantitative field. Nonetheless, useful feedback on the objectives was given as well.

The third interview helped a lot with better formulating the objectives and brought many useful insights that I adopted within my thesis. This interview helped a lot with the validation of my objectives and the related design requirements from a supervisory perspective.

The informal introductory talks before the interviews showed that currently banks are experimenting with machine learning, however they are quite risk-averse with using these techniques in application areas where there is a risk of discrimination.

The interviews all together helped with creating a more robust understanding of the practice of explicability evaluation of machine learning. Future research should be able to conduct more interviews and more empirical evidence with regards to the relevance and usefulness of the framework

4.4 Summary of chapter 4

Chapter 4 had the goal to answer two sub-questions, respectively 3 and 4: '*How can the objectives be transformed into design-requirements?*' and '*How can the design-requirements be transformed into a pragmatic prospective assessment framework?*'

For sub-question 3, the method by Van de Poel (2013) is used to formulate design-requirements that serve the higher objective ('*for the sake of this objective*'). 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. These design requirements are discussed on their rationale and interrelations.

With these design requirements, the subsequent step is answering sub-question 4. *The transformation into a pragmatic prospective assessment framework* continues with a morphological chart with (a) means for every requirement. 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).

The EAF needs to be adjusted to the context- and the explanation characteristics, aligned with the CRISP-DM lifecycle and complemented by user's guidelines that support the framework. Consequently, the framework consists of five sub-parts: *context characteristics*, *explanation characteristics*, *framework adjustments, assessment object assessment* and the *concluding section*. The EAF is a guiding table that needs to be filled according to the user's guidelines (appendix 3).

The conducted interviews for this thesis were carried out with a validation goal. According to these interviews with industry experts, the objectives as well as the design requirements are useful and important to include for the framework. Thus, it can be concluded that the evaluation phase can proceed with these objectives in order to test the quality of the framework, and that the validity of these aspects with regards to the industry is of a high level. In addition, the experts mentioned that such a framework is useful for the industry and this improves the validity with regards to the pragmatic aspect.

5. Chapter 5. Framework demonstration: case studies

This chapter encloses the demonstration of the designed assessment framework, which is the fourth phase of the DSRM (the demonstration phase). The framework will be applied to two cases that include produced explanations for a certain decision, which will be assessed. The first case entails an explanation by rule extraction and the second case an explanation by counterfactual explanation. By doing this, sub-question 5 will be answered: "*Is the framework pragmatic to use for assessment of explicability with a specific case?"*. We make use of the user's guidelines in appendix 3 to fill the EAF for the cases, in order to assess the explicability of the explanations and form a conclusion on the level of explicability of the machine learning system, whose decision is being explained.

5.1 Case 1 demonstration: Rule extraction explanation

5.1.1 Description case 1

The explanation to be assessed in the first case is an explanation by rule extraction of a support vector machine (SVM) (Martens, Huysmans, Setiono, Vanthienen, & Baesens, 2008). A support vector machine is one of the widely used machine learning techniques for classification that has the ability to include non-linear relations between variables, however, this technique is considered to have a high level of opaqueness. The application of rule extraction to mimic the SVM as close as possible helps to increase the explicability of this machine learning model and understand its logical internal workings. Rule extraction is a method to be used for the model explanation problem of chapter 2.3.1.1.

For our assessment, we use the rule extraction explanation that is produced by the RIPPER algorithm. This is an explanation of the machine learning system used for the classification of a German credit scoring data set from the UCI data repository, as documented in the paper (Martens et al., 2008). The classification is on the decision whether the applicant is a 'good'- or a 'bad' applicant.

The extracted rules introduce a problem: *the knowledge fusion problem*. This means that the output of applying RIPPER rule extraction could result in a rule that is unintuitive (a variable has a different effect on the outcome than expected) and therefore a reduction of the justification ability of the rules (Martens et al., 2008), which causes problems with implementation possibilities. Consequently, regarding the *evidence roles* to choose for the framework application, an explanation should include contrarian evidence and counter-evidence, if the decision is based on such evidence. Additionally, normal evidence and counter-evidence need to be included in the framework, so that the intuitive effects are covered. The

missing and exceptional (counter-) evidence are not included, since we do not have the resources to proof why certain evidence is expected by a layperson to have a larger effect (exceptional) than it has (the investigation of this is out-of-scope in this thesis), or that certain evidence is expected by a layperson to be there but is not included (missing) in the explanation. Future research and application of the framework by industry-experts (designers of the machine learning system) should reveal this.

The *intelligibility types* that are chosen to be included, in order to ensure the quality of the explanation, are the following three: 'why', 'how' and 'why not'. These intelligibility types are considered to be the most helpful in order to justify the decision that has been made. The 'input' information, different 'output' alternatives and 'what else' there is, is considered known by the explainee. 'Certainty', 'Control' and 'Situation' are considered irrelevant types to include since the goal is to justify the decision to a consumer: since the explanation needs to be based on the truth (requirement of soundness) giving probabilities on the justification of a decision is unnecessary (needs to be 100%), the application will not be investigated by the explainee and the situation is considered to be known by the applicant of the loan. The visualization intelligibility type is earlier mentioned as out-of-scope.

5.1.2 Framework application to case 1

Table 8: Filled EAF for case 1

1. Context characteristics	
Decision type	The machine learning system decides if the credit applicant is classified as good or bad (i.e. a high vs low probability of default)
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 '*good outstanding*' loans/less *bad outstanding* 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	"*Explanation capability towards the consumer is of crucial importance in a domain where the model needs to be validated before being implemented*"
Regulation(s)	N/a – no specific geographical area, but a more general use case
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. <u>Framework adjustments</u>		
Evidence roles selection (justification)		
	<u>Normal evidence</u>	**<u>Normal counter-evidence</u>**
	Exceptional evidence	Exceptional counter-evidence
	<u>Contrarian evidence</u>	**<u>Contrarian counter-evidence</u>**
	Missing evidence	Missing counter-evidence
Intelligibility types selection		
	Input	Output
	<u>Why</u>	**<u>How</u>**
	<u>Why not</u>	What if
	What else	Visualization
	Certainty	Control
	Situation	
4. <u>Assessment Object Assessment</u>		
Example rule set used to result in decision X (applicant is good/bad): *"if (Checking Account < 0DM) and (Housing = rent)* *then Applicant = Bad* *elseif (Checking Account < 0DM) and (Property = car or other) and* *(Present residence since ≤ 3y)* *then Applicant = Bad* *elseif (Checking Account < 0DM) and (Duration ≥ 30m)* *then Applicant = Bad* *elseif (Credit history = None taken/All paid back duly)* *then Applicant = Bad* *elseif (0 ≤ Checking Account < 200DM) and (Age ≤ 28) and* *(Purpose = new car)* *then Applicant = Bad* *else Applicant = Good"*		
Questions:	***Answers:***	
Justificatory explanation		
1.1 Does the explanation contain normal evidence?	No	
1.2 Does the explanation contain normal counter-evidence?	Yes, the *low amount at checking account, long duration, age < 28* 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, *housing = rent, property = car, purpose = new car, credit history = none taken/all paid back duly* are not expected to directly influence the decision negatively (it is not directly clear why this is the case), but it does.
Explanation towards a layperson	
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
Completeness	
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
Soundness	
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
Comprehensibility	
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

Conciseness	
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. <u>Concluding section</u>	
The explanation is insufficiently good, and therefore the ML system within this context and specific explanation is insufficiently explicable, because: - 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 In short: it lacks the layperson perspective, it is not complete enough, it is not comprehensible enough.	

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.

5.2 Case 2 demonstration: Counterfactual explanation

5.2.1 Description case 2

The paper by Grath et al. (2018) provides the second explanation to be assessed by using the EAF in this demonstration phase. This concerns another type of explanation, namely a *counterfactual explanation*. This explanation shows the minimum that input variables have to change in order to flip the decision from rejection to approval of the loan application (or the other way around). The paper describes a *counterfactual generation heuristic* that utilizes the loss function by Wachter et al. (2018) to provide the values that are required for the change.

The paper assesses the predictive power of four different classifiers for the classification challenge with the 'HELOC credit application dataset': *logistic regression, gradient boosting, support vector machine with linear kernel (SVC)* and *multi-layer perceptron* (the last three classifiers are considered black-box classifiers). The explanation generation algorithm is model-agnostic and is applied for all classifiers. The classification is on the decision whether a credit application is flagged as good or bad, which indicates that a consumer has paid all obligations vs a consumer that has been at least once 90 days past due within 24 months of opening a credit account.

The rationale for the *evidence roles* to include (normal/contrarian evidence/counter-evidence) is the same as the rationale for case 1: the knowledge fusion problem can occur with the black-box models, and the focus on a layperson as explainee (and the scope of this thesis) leads to the exclusion of the missing and exceptional (counter-)evidence. The *intelligibility types* to include are equal as well to the ones with the first case since the classification problem is the same here, and the context of the challenge did not change, so the type of explanation to give to a layperson should be the same.

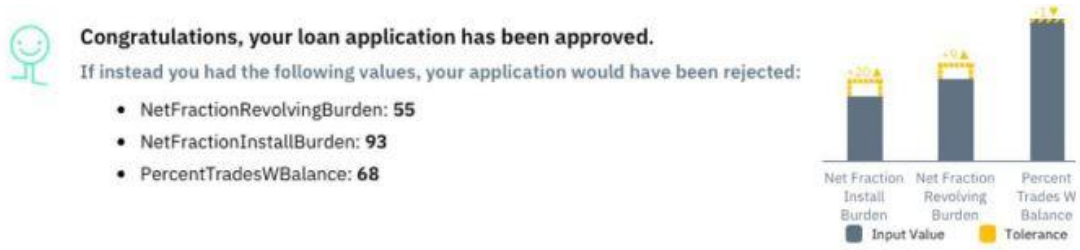
5.2.2 Framework application to case 2

Table 9: Filled EAF for case 1

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 '*good outstanding'* loans/less *bad outstanding* 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. <u>Explanation characteristics</u>		
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. <u>Framework adjustments</u>		
Evidence roles selection (justification)		
	<u>Normal evidence</u>	**<u>Normal counter-evidence</u>**
	Exceptional evidence	Exceptional counter-evidence
	<u>Contrarian evidence</u>	**<u>Contrarian counter-evidence</u>**
	Missing evidence	Missing counter-evidence
Intelligibility types selection		
	Input	Output
	<u>Why</u>	**<u>How</u>**
	<u>Why not</u>	What if
	What else	Visualization
	Certainty	Control
	Situation	
<u>Assessment Object Assessment</u>[5]		

[5] Explanation b (Grath et al., 2018) will be assessed, because of the more significant usefulness for real-world cases of explanations for rejected loans (which is the case with explanation b).



(a) Positive counterfactual explanation

(b) Counterfactual explanation

Questions:	***Answers:***
Justificatory explanation	
1.1 Does the explanation contain normal evidence?	Yes, *MSinceOldestTradeOpen* and *NumSatisfactoryTrades* 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, *NetFractionInstallBurden* and *NumBank2NatlTradesWHighUtilization* 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, *NumRevolvingTradesWBalance* is not expected to influence the decision strongly negative but it does.
Explanation towards a layperson	
2.1 What knowledge does the explainee need to have in order to understand the explanation?	The definitions and effects on the decision of: *NumRevolvingTradesWBalance, NetFractionInstallBurden, NumBank2NatlTradesWHighUtilization, NumSatisfactoryTrades* and *MSinceOldestTradeOpen*.

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.
Completeness	
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.
Soundness	
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.
Comprehensibility	
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
Conciseness	
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	
The explanation is insufficiently good, and therefore the ML system within this context and specific explanation is insufficiently explicable, because: - 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 In short: It lacks the layperson perspective, it is not complete enough, it is not comprehensible enough.	

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.

5.3 Summary of chapter 5

In this chapter, the Explicability Assessment Framework 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. The second one is a counterfactual explanation from a set of classification algorithms (including black-box algorithms such as support vector machine with linear kernel (SVC) and multi-layer perceptron). Herewith, the goal was to answer sub-question 5: “Is the framework pragmatic to use for assessment of explicability with a specific case?”. The user’s guidelines from appendix 3 are used to fill the framework. Both of the cases have the same explanation characteristics and slightly different context

characteristics. The adjustments to be made to the framework are the same for both cases. The explanations are different and produced by different techniques so the framework is independent of these aspects.

The first case study shows 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.

6. Chapter 6. Framework evaluation

This chapter describes the evaluation phase of the DSRM (phase 5), which entails the evaluation of the designed artifact on the objectives that are formulated in chapter 3. The goal is to answer sub-question 6: *"Does the demonstration show that the framework accomplished the objectives?"*.

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. 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.

Evaluation can take multiple forms, from logical and mathematical proof to any 'appropriate empirical evidence' (Peffers et al., 2008). Next to this evaluation on the objectives that are validated by interviewees from the industry, other evaluation techniques are very useful and should be conducted in future research, such as satisfaction surveys, interviews with the final users of the framework (the decision-makers with the development of the machine learning system). Time- and resource constraints within this thesis cause that these evaluations cannot be conducted in this thesis.

6.1 Evaluation of the framework on objectives

6.1.1 Provides guidance for the decision-makers on using the framework

This research aims to reduce the gap that exists between high-level principles of machine learning ethics and the usefulness for the decision-makers that design the machine learning systems, for the principle of explicability with as goal to improve accountability. To do so, the framework must be pragmatic and to enhance the clarity of this practical aspect of the framework, guidance for the users of the framework is required.

Taking the point-of-view of a decision-maker in the development lifecycle of a machine learning system, the question is whether this framework provides enough guidance so that the framework can be used in an effective way. The means for this guidance is the EAF user's guidelines from appendix 3. This has been used in the demonstration phase in order to assess the explanations in the two use cases and should all the time be attached to the EAF in order to ensure the right use 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).

The execution of these guidelines with the EAF on the use-cases resulted in the demonstration. The first case shows that the framework in combination with the user's guidelines provides enough guidance to fill out the framework and finalize with a conclusion regarding the explicability of the machine learning system with the specific context characteristics. Afterwards, recommendations can be formulated to move back to the business understanding phase in order to try to improve the explicability. The second case confirms this and increases the robustness of the framework since this is another case. It shows that different outcomes naturally lead to different recommendations for the business understanding step in order to enhance the explicability. Concluding, there is a sufficiently good guidance for decision-makers on using the framework.

The usefulness of the framework should, in any case, be validated in future research by end-users of the framework in the industry since the designer (the author of the thesis) of the framework is not pre-knowledge free. The reason for this is that all research for the design has been conducted by the author of this thesis, and naturally significantly increases the knowledge level. It could be possible that users of the framework do not directly possess this level, such that extra explanation and guidance is needed before the framework can successfully be used.

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

A framework is the most useful for decision-makers if it helps with the choice of whether a machine learning system that is developed to do a certain task, is explicable enough to be used for that task. Given that the other factors that play a role are sufficiently good (e.g. the performance of the machine learning system), the decision-makers can then move on to the deployment phase of the CRISP-DM lifecycle. The

framework must thus help to make a decision if the machine learning system is explicable enough such that the task can be delegated to it.

To do this, the framework must give an overview of the outcomes of the different assessment subsections that play a role in this choice. Further, it should show the users of the framework how the outcomes of the assessment can be synthesized to a yes or no on the question if they can proceed to the deployment phase.

The framework has a concluding section that summarizes the results of the assessment, after which a conclusion can be formulated on the question if the ML system is explicable enough, given the specific context that the framework has been adjusted to. The supporting user's guidelines give an explanation too on this section.

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. Thus, it can be concluded that the framework helps with the decision-making on whether a certain task or decision-making functionality can be delegated to the machine learning system.

6.1.3 Is able to prospectively assess explanations

In line with the Value Sensitive Design approach, a goal for the framework to accomplish was to be able to be used for prospective assessment. This is needed, because retrospective assessment is too late in the sense that, without explicability assessment before deployment of the machine learning system, problems with explicability and therefore accountability can already have occurred.

This section focuses on the question of whether it is possible to prospectively assess explanations with the EAF. An important note here is that we look at local ex-post explanations. This prospective view in combination with these types of explanations implies that we look at hypothetical decisions, and the demonstration phase has to show that this is possible to assess.

The demonstration is 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.

6.1.4 Is able to assess justificatory explanations

Justification of a decision has the most explanatory value for a layperson that is the subject of a decision. It is therefore needed that this framework can be adapted to- and used for justificatory explanations towards the explainee, for the accountability relationship between the bank and the credit applicant.

The framework provides two specific ways of adjusting the framework more to the justification towards a layperson aspect: the choices for *evidence roles* and for *intelligibility types*. With this goal in mind, the choices can be made accordingly such that the outcome is a framework specified for this. The guidelines provide a description of all the options and the user should make a supported choice why the specific choice has been made.

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.

Future empirical research should be devoted to what options are considered the most relevant for a layperson to include in an explanation for the justification goal.

6.1.5 Is able to assess consumer-level explanations towards a layperson

This objective ensures the ability of the framework to assess explanations that are meant for explainees with basic knowledge, limited understanding of the context and little education. The context of credit applications requires this since all of society can apply for a loan and they should be able to receive an explanation that is understandable.

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 and the comprehensibility and conciseness sections (chapter 6.1.8 and 6.1.9 will go into detail with the evaluation of comprehensibility and conciseness). 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.

6.1.6 Is able to assess the completeness of explanations

The completeness is the first of the 'four principles for good explanations' to be assessed. The framework makes use of the intelligibility types formulated by Lim & Dey (2009). Kulesza et al. (2013) already use a method to assess the completeness of an explanation based on the appearance of different intelligibility types in the explanation and we adopt this method.

The demonstration shows that first, the intelligibility types 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.

6.1.7 Is able to assess the soundness of explanations

Kulesza et al. (2013, p.2) describe soundness as *"the extent to which each component of an explanation's content is truthful in describing the underlying system"*. This implies that there is a scale within this goodness principle, and it is context-dependent what level of soundness is required for an explanation. The framework should outline the different levels and be able to be used to assess the explanation on the level of soundness that is present there.

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.

We can conclude that by using the framework we can validate the soundness of the consumer-level explanation 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 sensure that the explanation is fully sound on the required level. The assessment of explanations on this level is out-of-scope in this thesis and requires research in order to design a framework to accomplish this (or extend this framework with this point of view).

Further research should be conducted on the following question: what is the minimum level of soundness that an explanation needs in what context (in addition to the floor minimum of no false or incorrect statements)?

6.1.8 Is able to assess the comprehensibility of explanations

While not underestimating the other three principles for explanation goodness, comprehensibility is clearly very important for an explanation towards a layperson. The explainee needs to be able to understand an explanation that has been given to him or her. The question is: does the framework contain the handles for the decision-makers to be able to assess the comprehensibility of an explanation?

The framework contains multiple aspects that help to address this goal: the assessment of the *narrative format, contextual supportive text opposed to singular facts/data points, understandability of the used language* (related to the comprehensibility principle) and *the logical relation between the decision, features, roles, and effects*. The guidelines explain the different aspects and support why they are important.

Within the demonstration, we can see that the assessment of these aspects in both of the explanations results in recommendations for improvement. We conclude that the EAF thus can be used for assessment of the comprehensibility principle of an explanation. 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.

6.1.9 Is able to assess the conciseness of explanations

Lastly, the conciseness of an explanation is a principle that needs to be assessed due to the fact that more concise explanations are often considered easier to understand, although there is a tension with completeness and the goal to reach an optimum: a too-short explanation is hard to understand and not complete enough, but a too-long explanation is hard to understand as well and less comprehensible (the explainees lose their attention).

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.

6.2 Limitations

In this thesis, 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 thesis 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.

6.3 Summary of chapter 6

This chapter focused on the evaluation of the framework and the demonstration of the framework based on the achievement of the objectives. Sub-question 6 will be answered by means of this: '*Does the demonstration show that the framework accomplished the objectives?*'. The main points of the evaluation by objective are the following:

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.

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.

C. *Is able to prospectively assess explanations*

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

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.

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

Yes, the framework can be adjusted towards this perspective and thus assess the explanations on this perspective, however, future research should be devoted to empirically investigate what knowledge level a layperson has.

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.

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. To fully do this, the machine-level and business-level soundness should be

ensured. In addition, future research should be conducted to investigate what the minimum level of soundness is that an explanation needs in a specific context.

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.

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.

Furthermore, this thesis and the designed artifact have some limitations that need attention. First, the framework is a step in the right direction, but it has several prerequisites and covers a modest part of the full explicability spectrum. Second, the evidence roles and intelligibility types are currently chosen and need validation from the real-world. Third, a threshold is needed for soundness and conciseness in order to fully assess these aspects. Four, the proposed users of the framework have not yet worked with the EAF itself and this could increase the validity of the pragmatic aspect. Five, the knowledge level of a layperson should be empirically researched to create a more hands-on possibility to assess the comprehensibility. If time constraints had allowed, a second iteration could be made with empirical validation elements, such that the framework could be improved on this aspect.

7. Chapter 7. Conclusions

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 gathered and discussed here. Moreover, this chapter is the largest part that contributes to the communication phase of the DSRM (phase 6). First, chapter 7.1 includes the main findings in sequential order, including the answers of the sub-questions, and finalizing with an answer to the main research question. Chapter 7.2 elaborates on the interpretation of these main findings, including importance- and utility of the framework, and the generalizability of the findings. Further, chapter 7.3 discusses the recommendations for future research, which inter alia entails ways how to cope with the earlier mentioned limitations. Next, chapter 7.4 discusses how this thesis relates to the MSc. program. Complex Systems Engineering and Management (CoSEM) program of Delft University of Technology (TU Delft), and the scientific and societal contribution of this research (chapter 7.5). Lastly, this thesis concludes with chapter 7.6 with a more general reflection on artificial intelligence and explicability by the author of this thesis.

7.1 Main findings

This thesis addressed the problem that it is unclear for decision-makers within the development of a machine learning system for credit underwriting how to structurally evaluate explicability of such a system. More extensive, the answer at sub-question 1 (what is the problem that the objectives should solve?) is:

"It is unclear for decision-makers within the development of machine learning systems how to structurally evaluate explicability in credit underwriting cases. To enhance explicability, a pragmatic qualitative framework is needed, useful for prospective assessment of machine learning systems from the point of view of the explicability"

The goal of this research was to design a pragmatic qualitative framework for prospective assessment of a use case. By reducing this ambiguity, it becomes easier for developers and decision-makers (the users of the framework) to be sure if a machine learning system needs improvement in terms of explicability, which is an important step towards compliance with the GDPR and CCD and positively contributes to the ethical need for explanation of consumers. To accomplish this objective, the following research question was drafted: *"How can decision-makers prospectively assess machine learning applications within credit underwriting from the point of view of explicability?"*

First, an extensive literature overview has been created by performing desk-research. Machine learning used in the financial services industry is introduced, after which an overview of explicability and machine learning is given. Further, different types of explanations have been discussed moving slowly towards the scope of this thesis within the outlined explanation research field, and finally, the explanation goodness principles that play a large role within the assessment. Lastly, the Value Sensitive Design (VSD) approach and the Design Science Research Methodology (DSRM) are elaborated upon, that both collectively form the methodology and process structure of this thesis.

The DSRM consists of six sequential phases that are performed in this thesis: *problem identification, objectives definition, design & development, demonstration, evaluation, and communication*. In chapter 1 the problem is identified, after which the *design objectives* are formulated in chapter 3. This is directly the answer of sub-question 2: *What are the objectives for the assessment framework?*

The framework:

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

Design requirements are derived from these objectives, that are empirically validated by industry experts through semi-structured interviews, to accomplish them. The answer of sub-question 3 (*How can the objectives be transformed into design-requirements?)* is the following: by using the method that is mentioned by van de Poel (2013) functional design-requirements and constraints are formulated that 'satisfy the upper norm' of one of the former objectives. These design requirements satisfy a 'for the sake of' relation towards this objective.

With these design requirements, the more creative step framework development is taken. Here the answer of sub-question 4 (*How can the design-requirements be transformed into a pragmatic prospective assessment framework?*) is formulated. Means are described in a function-means diagram to meet the requirements and the means are synthesized into the actual artifact, the assessment framework. Figure 14 shows the framework with the interrelations that influence the framework.

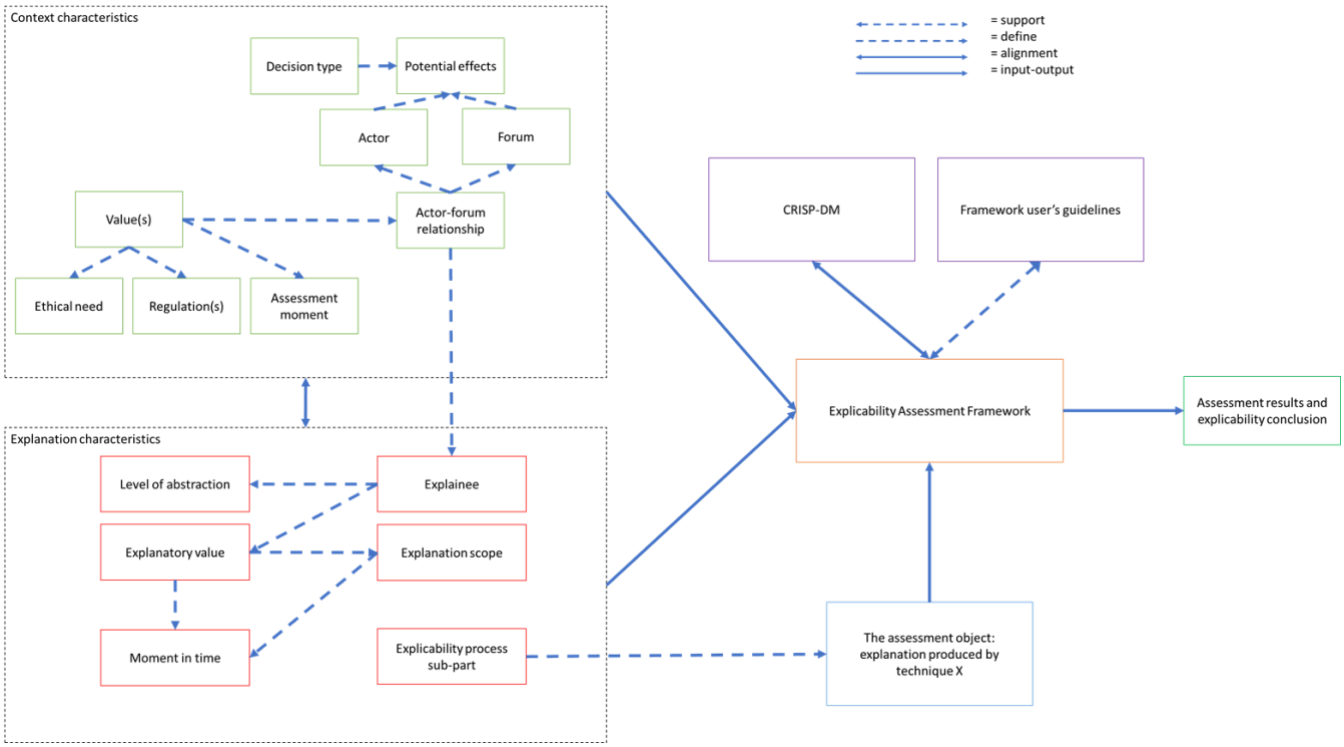


Figure 13: Explicability Assessment Framework and relationships

The framework consists of two large artifacts: the actual framework, or the *Explicability Assessment Framework (EAF),* and the supporting user's guidelines. The EAF has four subsections that the user has to fill in order to ultimately formulate a conclusion that can be documented in the concluding section: *the context characteristics, the explanation characteristics, the framework adjustments,* and *the assessment object assessment.* The first one is a section where the context characteristics can be described*: the decision type, the value(s), the actor-forum relationship, the forum, the actor, the potential effects, the ethical need* and *the regulation(s)*. These characteristics induce (partly) the choices in the second section on the explanation characteristics: *the explainee, level of abstraction, explanatory value, explanation scope, moment in time* and *explicability process sub-part*. The scope of the designed framework in this thesis is the following: *explanation product (explicability process sub-part), prospective (assessment moment), justification (explanatory value), local (explanation scope), ex-post (moment in time), consumer-*

level (level of abstraction) and *layperson (explainee)*. The framework adjustments (what *evidence roles* and what *intelligibility types* to include) are derived from the first and second section. The demonstrated framework makes use of *the normal-* and *contrarian evidence* and *counter-evidence roles,* and *the why-, how-* and *why not-types.* The fourth section is the actual assessment of the explanation object on the following aspects: *justification, layperson-perspective, completeness, soundness, comprehensibility* and *conciseness*. Afterwards, the concluding section can take place, based on the performed assessment.

The user's guidelines consist of three main sections that support the EAF. The first section is an important statement regarding the prerequisites for using the framework and the second section includes an introduction to the guidelines. The third section is the actual user's guidelines that systematically move along the AEF, starting with the AEF itself. It describes where and how the framework fits in the CRISP-DM lifecycle and continues with a structured description of the task to be performed, including additional explanation and/or an example. It finalizes with a description of the tasks and interpretation of the concluding section of the EAF.

The framework is demonstrated with the application on two different use cases where a machine learning technique is used for credit underwriting, and an explanation technique is chosen to formulate an explanation that is being assessed. The guiding sub-question for this phase is sub-question 5: *Is the framework pragmatic to use for assessment of explicability with a specific case?* In both of the cases the explanation lacks on specific points, and this is succeeded by recommendations for the users of the framework to improve the explanation. One can conclude from the conducted demonstration that the framework, in combination with the guidelines, is pragmatic and provides handles and guidelines in order to help the decision-makers and provide them with next steps to take with regards to the outcome of the assessment.

The evaluation phase takes place hereafter, in which the demonstration and framework are evaluated on the extent to which the objectives are reached. In this chapter, the last sub-question is answered: *Does the demonstration show that the framework accomplished the objectives?* The evaluation phase gives rise to the answer that most objectives are accomplished, and some objectives are partly accomplished but need more research in order to fully fulfill these goals. In this section, limitations with regards to the objectives are identified and discussed, next to the evaluation of the execution of the demonstration. Recommendations for future research to cope with these limitations will be discussed further on in this chapter.

Shortly summarizing all these main findings, 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 (explainee). 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.

7.2 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 that are formulated by Floridi et al. (2018, p. 13-14), 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."*

This thesis contributes as well to two of the "*seven essentials for achieving trustworthy AI*", as defined by the European Commission High-Level Expert group on AI (European Commission, 2019):

- ***"Diversity, non-discrimination and fairness:*** *AI systems should consider the whole range of human abilities, skills and requirements, and ensure accessibility."*
- **"*Accountability:*** *Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes."*

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. For instance, since it can be argued that auditors represent the controllers for the rules that follow from our society, this is a good first step.

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. One could think of automated legal decisions and even the explicability of autonomous systems in self-driving cars. With both of these cases accountability is a highly valuable principle that should be committed to.

Despite that 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.

7.3 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 6.2). 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.

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 certain 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 (2017) 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.

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.

Future research could as well be conducted to investigate to what extent this framework could have been developed differently with regards to using the axiomatic design approach (a design optimization approach), as discussed in chapter 4.

Lastly, the conducted interviews (appendix 2) 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 of the second interview (appendix 2.2), 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 within the introductory informal talks with the interviewees 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.

7.4 Scientific and societal contribution

Scientific contribution

The research gap that this thesis aimed to fill is the following: the scientific literature lacks research on how reduce the gap between a high-level principle like explicability and the operational level of implementing such a value in the development cycle of a machine learning system. Thereupon, prospective assessment of a case on this principle required research. Besides that, most literature takes a mono-disciplinary viewpoint, where explicability assessment requires a multidisciplinary perspective. Furthermore, the thesis should add a multi-organizational guiding assessment framework that is aligned with a machine learning system development lifecycle (a design-oriented perspective) to the literature.

This thesis has addressed this gap by the derivation and formulation of nine context characteristics and six explanation characteristics that influence the assessment framework. It shows that by describing and analyzing the context, the explanation characteristics can be shaped, and these collectively can adapt the assessment framework that can be used to evaluate explicability. A multidisciplinary view was taken and regulatory, social, ethical, business, procedural and technological aspects are all synthesized in the preparations and the actual assessment framework. Further, user's guidelines have been drafted that support the use of the assessment framework to optimize the pragmatic aspect of the framework as much

as possible. In addition, the CRISP-DM lifecycle was analyzed and an assessment process has been described (how to use the EAF) in order to align the framework with an inter-organizational development lifecycle model.

The thesis has provided us a tool to systematically evaluate explanations. Moreover, it gives us insight into the interrelation of the context characteristics and how this shapes the way the assessment needs to take place.

To the best of the author's knowledge, no literature has directly described the interrelations of the influencing factors for the assessment or has dealt with the transformation of the high-level principle of explicability towards the level of actual assessment of a machine learning system use case. Therefore, these can be considered as valuable contributions to the machine learning ethics literature.

Societal contribution

The problem that this thesis was dealing with was: it is unclear for decision-makers within the development lifecycle how to structurally evaluate explicability of machine learning systems in credit underwriting cases. This problem has been addressed by the provision of the EAF and the user's guidelines together. The EAF gives the decision-makers the tool for the assessment and the user's guidelines show them how to use the tool. It helps to solve a twofold of issues. First, the GDPR and CCD prescribe that automated decisions need to be able to be explained towards the consumer, although it is arguable if the right to explanation exists within the GDPR. Nonetheless, the regulations show the idea and direction of the European government and this could indicate future regulations as well. Second, the issue of an ethical need for explanation of automated decisions can largely be appeased, when explicability is improved by using the framework to evaluate it and improve the machine learning system on explicability afterwards.

Moreover, the unclarity for decision-makers is decreased and this helps with the improvement of compliance with regulations and to serve the ethical need that exists in society.

7.5 Link of study and thesis

CoSEM aims to teach its students to be able to design socio-technical systems. We talk here about a broader field than technological innovation and include the governance and management of the technological artifact. This interdisciplinary field requires an in-depth technical understanding of the innovation and a strong system-perspective to cope with regulations and societal challenges. The courses '*Complex Systems Engineering* and *CoSEM research challenges'* were the master's courses that especially inspired this thesis subject and type.

There is a clear link between this thesis and the master's program for a couple of reasons. First, the challenge in this thesis is multidisciplinary pre-eminently: technological complexity with machine learning innovation, ethical complexity with the implications of the technological artifact, regulatory complexity to cope with this new innovation and business complexity due to business processes and needs. Second, the design challenge of developing an artifact that is able to deal with the ethical problem that could arise with machine learning is evidently present in this thesis. Third, this thesis uses a systematical and structured method (DSRM) from the program to evaluate and synthesize information, and ultimately develop a framework that is based on this. This thesis is a perfect example of a status-quo interdisciplinary challenge that multidisciplinary CoSEM students are arguably par excellence capable of dealing with.

7.6 Reflection on artificial intelligence ethics

Having finalized the research project for this master's thesis, this section takes a broader view and reflects on artificial intelligence ethics in general. Concerning this subject, what are prospects for the future? How should commercial companies act with regards to AI ethics? What about academic- and scientific organizations? And what are future steps that (semi-)public organizations should take on? After a significant period of time immersing in the scientific field of machine learning explicability, this paragraph will discuss the personal view of the author.

Considering how much scientific- and professional research there currently is being conducted within the field of artificial intelligence, future progressions are looking bright, with the self-evident condition that risks are minimized as much as possible. This research was first focused on all the advantages that the use of artificial intelligence techniques has, and can have, but with the evolution of this research more issues came to the surface, that consist of a significant amount of ethical risks. It is now recognized that significant effort must be made in order to ultimately ensure ethical- deployment and use of artificial intelligence systems. So, what can we expect?

From the reactive perspective, a risk to look out for is that organizations can't see the forest for the trees since this topic covers a wide range of values (e.g. from privacy to sustainability, from explicability to human autonomy, and from social inclusion to accountability). There are tensions or even tradeoffs between values, and one could be too expectant to make decisions, due to the fear to forget or neglect an important other value. Therefore, for oversight organizations, one of the main challenges that need to be dealt with in the near future is to hand guidance to these organizations by developing a clear strategy towards the reduction of AI ethics risks. Governmental guidelines could be very helpful and especially

transparency about the future steps of the government is important: this reduces uncertainty and consequently the risk of bad investments for organizations. This strategy should further include some initial steps for organizations to take when developing and/or using AI systems, concerning AI ethics.

On the more proactive side, one should look out for *ethics washing:* the effort that seems to improve AI ethics within the company, but with a commercial purpose as rationale, instead of the common good (e.g. an ethics advisory board opposed to real regulations). If this improves AI ethics, it does seem like not much of a problem, however, an issue could evolve when the commercial goal is reached and ethical principles are not thoroughly incorporated in the company. Herewith, a severe risk exists that the company (unknowingly) returns to the worse old ethical situation while having the commercial benefit of the temporary better ethical situation. Transparency about the real actions that the company takes in relation to the improvement of AI ethics could help to avoid these situations.

Further, when the government opens up about their strategy and severe real effort is made to reduce AI ethics risks, companies can move forward and invest in research, development and partnerships in order to deploy ethical and trustworthy AI. It is advisable to start with governmental guidelines as starting point, and explore the potential areas that the business has ethical risks that need to be coped with. Next, a company can proceed with an initial explorative risk assessment and evaluating possibilities to reduce these risks.

Lastly, (semi-) public organizations have an important role to facilitate partnerships and AI ethics *ecosystems*. There are different ways to pursue this. One could approach potential companies for a first meeting on a specific topic that could lead to more frequent meetings between companies, to work on and share knowledge of this topic, in order to increase the potential of synergistic effects: bilateral knowledge and/or value increase. Another option is to (partly) fund projects that have the goal to improve AI ethics within the company. Companies tend to be waiting for the government to take the first steps, and in order to improve AI ethics all along the industry it is important for them to be involved. Furthermore, public organizations can ask companies from a lot of different disciplines to collaborate and this could help with solving the interdisciplinary problems that this subject possesses.

8. Chapter 8. References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Algorithmic Accountability Act of 2019, H.R.2231, 116th Cong. (2019).

Angwin, J., Varner, M., & Tobin, A. (2017, September 14). Facebook Enabled Advertisers to Reach 'Jew Haters.' Retrieved March 13, 2019, from ProPublica website: https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

Altman, A. (2016). Discrimination. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2016). Retrieved from https://plato.stanford.edu/archives/win2016/entries/discrimination/

Balasubramanian, R., Libarikian, A., & McElhaney, D. (2018, April). Insurance 2030—The impact of AI on the future of insurance | McKinsey. Retrieved March 13, 2019, from https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance

Bartlett, R. P., Morse, A., Stanton, R., & Wallace, N. (2017). *Consumer Lending Discrimination in the FinTech Era* [Working Paper]. Retrieved from UC Berkeley Public Law website: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063448

Bennink, K. (2017). *Do traditional banks need a Chief AI Officer?: An explorative research project that aims to evaluate the appointment of a Chief AI Officer to overcome challenges that arise when traditional banks adopt AI technologies*. Retrieved from http://resolver.tudelft.nl/uuid:081758b4-a227-4e18-9950-97da9d49396c

Berthold, M., & Hand, D. J. (2007). *Intelligent Data Analysis: An Introduction* (2nd ed.). Berlin, Heidelberg: Springer Berlin Heidelberg.

Biran, O., & McKeown, K. (2014). Justification narratives for individual classifications. *AutoML Workshop*, *32*, 7. Retrieved from http://www.cs.columbia.edu/nlp/papers/2014/justification_automl_2014.pdf

Boillet, J. (2018, April 1). Why AI is both a risk and a way to manage risk. Retrieved March 13, 2019, from https://www.ey.com/en_gl/assurance/why-ai-is-both-a-risk-and-a-way-to-manage-risk

Borgo, R., Cashmore, M., & Magazzeni, D. (2018). Towards Providing Explanations for AI Planner Decisions. *Proceedings of IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)*, 11-17. Retrieved from http://arxiv.org/abs/1810.06338

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, *13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Buss, S., & Westlund, A. (2018). Personal Autonomy. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2018). Retrieved from https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Bygrave, L. A. (2001). Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling. *Computer Law & Secuirty Review*, *17*(1), 17–24. https://doi.org/10.1016/S0267-3649(01)00104-2

Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). *Plan explanations as model reconciliation: Moving beyond explanation as soliloquy*. 156–163. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85031946471&partnerID=40&md5=8feca0b28d481518c2eaf3dcb07866f7

Chakraborti, Tathagata, Kulkarni, A., Sreedharan, S., Smith, D. E., & Kambhampati, S. (2018). Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. *In Proceedings of the International Conference on Automated Planning and Scheduling (Vol. 29, No. 1, pp. 86-96)*. Retrieved from https://wvvw.aaai.org/ojs/index.php/ICAPS/article/download/3463/3331/

Crawford, K. (2017). *The Trouble with Bias—NIPS 2017 Keynote*. Retrieved from https://www.youtube.com/watch?v=fMym_BKWQzk

Cui, X., Lee, J. M., & Hsieh, J. P.-A. (2019). An Integrative 3C evaluation framework for Explainable Artificial Intelligence. *AMCIS 2019 Proceedings*, *25*, 1–10. Retrieved from https://aisel.aisnet.org/amcis2019/ai_semantic_for_intelligent_info_systems/ai_semantic_for_intelligent_info_systems/10

Davis, J., & Nathan, L. P. (2015). Value Sensitive Design: Applications, Adaptations, and Critiques. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 11–40). https://doi.org/10.1007/978-94-007-6970-0_3

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Dignum, V. (2017). Responsible Autonomy. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4698–4704. https://doi.org/10.24963/ijcai.2017/655

Dignum, V., Lopez-Sanchez, M., Micalizio, R., Pavón, J., Slavkovik, M., Smakman, M., … Kließ, M. S. (2018). Ethics by Design: Necessity or Curse? *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, 60–66. https://doi.org/10.1145/3278721.3278745

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. Pub. L. No. 32008L0048, 133 (2008).

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv:1702.08608 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1702.08608

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., … Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal*, *18*(7). https://doi.org/10.2139/ssrn.3064761

Dym, C. L., Little, P., & Orwin, E. J. (2014). *Engineering design: A project-based introduction* (4. ed). New York: Wiley.

European Central Bank. (2018). *Financial Stability Review, November 2018* (p. 170). Retrieved from https://www.ecb.europa.eu/pub/pdf/fsr/ecb.fsr201811.en.pdf

European Commission. (2018, June 14). High-Level Expert Group on Artificial Intelligence [Text]. Retrieved April 19, 2019, from Digital Single Market—European Commission website: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

European Commission. (2019, August 4). Artificial intelligence: Commission takes forward its work on ethics guidelines. Retrieved October 14, 2019, from https://europa.eu/rapid/press-release_IP-19-1893_en.htm

Faulconbridge, R. I., & Ryan, M. J. (2014). *Systems Engineering Practice*. Canberra, Australia: Argos Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fox, M., Long, D., & Magazzeni, D. (2017). Explainable Planning. *ArXiv:1709.10256 [Cs]*. Retrieved from http://arxiv.org/abs/1709.10256

Freitas, A. A. (2001). Understanding the Crucial Role of Attribute Interaction in Data Mining. *Artificial Intelligence Review*, *16*(3), 177–199. https://doi.org/10.1023/A:1011996210207

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). https://doi.org/10.1007/978-94-007-7844-3_4

Frosst, N., & Hinton, G. (2017). Distilling a Neural Network Into a Soft Decision Tree. *ArXiv:1711.09784 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1711.09784

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89. https://doi.org/10.1109/DSAA.2018.00018

Grath, R. M., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018). Interpretable Credit Application Predictions With Counterfactual Explanations. *NIPS 2018 workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, Dec 2018, Montreal, Canada.* Retrieved from https://hal.inria.fr/hal-01934915

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. *ArXiv:1805.10820 [Cs]*. Retrieved from http://arxiv.org/abs/1805.10820

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D. (2017, November). *Explainable Artificial Intelligence (XAI)*. Retrieved from https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Hajian, S., & Domingo-Ferrer, J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

Hansson, M. (2019). Artificial Intelligence for Humans. *Think*, *1*, 20.

Havard, C. (2011). "On the Take": The Black Box of Credit Scoring and Mortgage Discrimination. *Boston University Public Interest Law Journal*, *20*(2), 241–287. http://dx.doi.org/10.2139/ssrn.1710063

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For "Explainable Ai." Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62(1), 197–201. https://doi.org/10.1177/1541931218621047

Holzer, H. J., & Neumark, D. (2006). Affirmative action: What do we know? *Journal of Policy Analysis and Management*, *25*(2), 463–490. https://doi.org/10.1002/pam.20181

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *ArXiv:1712.09923 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1712.09923

IBM. (2014, October 24). CRISP-DM Help Overview. Retrieved July 2, 2019, from undefined

Ince, H., & Aktan, B. (2009). A Comparison of Data Mining Techniques for Credit Scoring in Banking: A Managerial Perspective. *Journal of Business Economics and Management*, *10*(3), 233–240. https://doi.org/10.3846/1611-1699.2009.10.233-240

International Covenant on Civil and Political Rights. (1966). *United Nations, Treaty Series*, *999*, 171–185.

Kannetzky, F. (2002). Expressibility, Explicability, and Taxonomy. In G. Grewendorf & G. Meggle (Eds.), *Speech Acts, Mind, and Social Reality: Discussions with John R. Searle* (pp. 65–82). https://doi.org/10.1007/978-94-010-0589-0_5

Koene, A., Clifton, C. W., Hatada, Y., Webb, H., Patel, M., Machado, C., … Scientific Foresight Unit. (2019). A governance framework for algorithmic accountability and transparency: Study. Retrieved from http://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (5th ed.). Retrieved from https://books.google.nl/books?id=xYRDAAAAQBAJ

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, 126–137. https://doi.org/10.1145/2678025.2701399

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

Kulkarni, A., Zha, Y., Chakraborti, T., Vadlamudi, S. G., Zhang, Y., & Kambhampati, S. (2019). Explicable Planning As Minimizing Distance from Expected Behavior. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2075–2077. Retrieved from http://dl.acm.org/citation.cfm?id=3306127.3332015

Lewis, D. (1986). Causal Explanation. In *Philosophical Papers*: *Vol. 2*. *Causation* (pp. 214–240).

Lim, B. Y., & Dey, A. K. (2009). Assessing Demand for Intelligibility in Context-aware Applications. *Proceedings of the 11th International Conference on Ubiquitous Computing*, 195–204. https://doi.org/10.1145/1620545.1620576

Lipton, Z. C. (2018). The Mythos of Model Interpretability. Queue, 16(3), 30:31–30:57. https://doi.org/10.1145/3236386.3241340

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2012). *Explanation and Abductive Inference*. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

Lui, A., & Lamb, G. W. (2018). Artificial intelligence and augmented intelligence collaboration: Regaining trust and confidence in the financial sector. *Information & Communications Technology Law*, *27*(3), 267–283. https://doi.org/10.1080/13600834.2018.1488659

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1705.07874

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., & Baesens, B. (2008). Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. In J. Diederich (Ed.), *Rule Extraction from Support Vector Machines* (pp. 33–63). https://doi.org/10.1007/978-3-540-75390-2_2

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Retrieved from http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McLeod, C. (2015). Trust. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2015). Retrieved from https://plato.stanford.edu/archives/fall2015/entriesrust/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Ministerie van Financiën. Regeling eed of belofte financiële sector 2015. BWBR0036152 FM 2014/1237 M § (2015).

Mittal, V., & Gupta, S. (n.d.). *KYC automation using artificial intelligence (AI)*. Retrieved from https://www.ey.com/Publication/vwLUAssets/ey-kyc-automation-using-ai/$FILE/ey-kyc-automation-using-ai.pdf

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 279–288. https://doi.org/10.1145/3287560.3287574

Mok, L., & Hyysalo, S. (2018). Designing for energy transition through Value Sensitive Design. Design Studies, 54, 162–183. https://doi.org/10.1016/j.destud.2017.09.006

Mueller, S. T., Hoffman, R. R., Clancey, W. J., Emery, A. K., & Klein, G. (2019). Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI (No. TA-2_02/19). Retrieved from Florida Institute for Human and Machine Cognition Pensacola United States website: https://apps.dtic.mil/docs/citations/AD1073994

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, *2*(1), 25–42. https://doi.org/10.1007/BF02639315

Papageorgiou, E. I., Spyridonos, P. P., Stylios, C. D., Ravazoula, P., Groumpos, P. P., & Nikiforidis, G. N. (2006). Advanced soft computing diagnosis method for tumour grading. *Artificial Intelligence in Medicine*, *36*(1), 59–70. https://doi.org/10.1016/j.artmed.2005.04.001

Partington, A., & Pichler, A. (2013). *Future Identity Banking*. Retrieved from https://www.accenture.com/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Dualpub_9/Accenture-Future-Identity-Banking.pdf

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware Data Mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568. https://doi.org/10.1145/1401890.1401959

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 45–77.

Piatesky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved July 28, 2019, from https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in Explainable AI. *ArXiv:1810.00184 [Cs]*. Retrieved from http://arxiv.org/abs/1810.00184

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). Pub. L. No. 32016R0679, 119 OJ L (2016).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*. https://doi.org/10.1007/s11023-019-09509-3

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, New Jersey, NJ, USA: Pearson Education, Inc.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, *1*, 1–10.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *44*(1.2), 206–226. https://doi.org/10.1147/rd.441.0206

Seegebarth, B., Müller, F., Schattenberg, B., & Biundo-Stephan, S. (2012). Making Hybrid Plans More Clear to Human Users—A Formal Approach for Generating Sound Explanations. *ICAPS*.

Sengupta, S., Chakraborti, T., Sreedharan, S., Vadlamudi, S. G., & Kambhampati, S. (2017). *RADAR - A Proactive Decision Support system for human-in-the-loop planning*. *FS-17-01-FS-17-05*, 269–276. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044462904&partnerID=40&md5=d20d192096e13d0db07c7ef7af9d2b3f

Setiono, R., & Liu, H. (1995). Understanding Neural Networks via Rule Extraction. *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, 480–485. Retrieved from http://dl.acm.org/citation.cfm?id=1625855.1625918

Sheh, R., & Monteath, I. (2018). Defining Explainable AI for Requirements Analysis. *KI - Künstliche Intelligenz*, *32*(4), 261–266. https://doi.org/10.1007/s13218-018-0559-3

Simpson, T. W. (2012). What Is Trust? Pacific Philosophical Quarterly, 93(4), 550–569. https://doi.org/10.1111/j.1468-0114.2012.01438.x

Streefkerk, R. (2019, April 25). Transcribing an interview in 5 steps. Retrieved October 16, 2019, from Scribbr website: https://www.scribbr.com/methodology/transcribe-interview/

Sreedharan, S., Chakraborti, T., & Kambhampati, S. (2018). *Handling model uncertainty and multiplicity in explanations via model reconciliation*. *2018-June*, 518–526. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054768722&partnerID=40&md5=1fe2a224d286832567c4d55754397c87

Sreedharan, Sarath, Chakraborti, T., Muise, C., & Kambhampati, S. (2019). Planning with Explanatory Actions: A Joint Approach to Plan Explicability and Explanations in Human-Aware Planning. *ArXiv:1903.07269 [Cs]*. Retrieved from http://arxiv.org/abs/1903.07269

Sreedharan, S., Chakraborti, T., & Kambhampati, S. (2017). Balancing Explicability and Explanation in Human-Aware Planning. 2017 AAAI Fall Symposium Series. Presented at the 2017 AAAI Fall Symposium Series. Retrieved from https://www.aaai.org/ocs/index.php/FSS/FSS17/paper/view/16030

Suh, N. P. (1998). Axiomatic Design Theory for Systems. Research in Engineering Design, 10(4), 189–209.

Taebi, B., Correljé, A., Cuppen, E., Dignum, M., & Pesch, U. (2014). Responsible innovation as an endorsement of public values: The need for interdisciplinary research. *Journal of Responsible Innovation*, *1*(1), 118–124. https://doi.org/10.1080/23299460.2014.882072

Trippi, R. R., & Turban, E. (Eds.). (1992). *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*. New York, NY, USA: McGraw-Hill, Inc.

Turing, A. M. (1950). Computer Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vaishnavi, V., Kuechler, W., & Petter, S. (2017, December 20). *Design Science Research in Information Systems*. Retrieved from http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf

van de Poel, I. (2013). Translating Values into Design Requirements. In D. P. Michelfelder, N. McCarthy, & D. E. Goldberg (Eds.), *Philosophy and Engineering: Reflections on Practice, Principles and Process* (Vol. 15, pp. 253–266). https://doi.org/10.1007/978-94-007-7762-0_20

van den Hoven, J., Lokhorst, G.-J., & Van de Poel, I. (2012). Engineering and the Problem of Moral Overload. *Science and Engineering Ethics*, *18*(1), 143–155. https://doi.org/10.1007/s11948-011-9277-z

Velleman, J. D. (2003). Narrative Explanation. *The Philosophical Review*, *112*(1), 1–25. Retrieved from JSTOR.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, *31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Weld, D. S., & Bansal, G. (2019). The Challenge of Crafting Intelligible Intelligence. Commun. ACM, 62(6), 70–79. https://doi.org/10.1145/3282486

Weller, A. (2019). Transparency: Motivations and Challenges. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Muller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 23–40). https://doi.org/10.1007/978-3-030-28954-6_2

Williams, T., Szafir, D., Chakraborti, T., & Ben Amor, H. (2018). *Virtual, Augmented, and Mixed Reality for Human-Robot Interaction*. 403–404. https://doi.org/10.1145/3173386.3173561

Winfield, A. (2019). Ethical standards in robotics and AI. *Nature Electronics*, *2*(2), 46. https://doi.org/10.1038/s41928-019-0213-6

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, 29–39. Manchester, UK: Practical Application Company.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In 8th International Conference on Evaluation and Assessment in Software Engineering (EASE 2014, 321–330. ACM.

ZestFinance. (n.d.). ZAML® Credit and Risk Modeling Solutions | ZestFinance. Retrieved March 5, 2019, from https://www.zestfinance.com/zaml

Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H. H., & Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 1313–1320. https://doi.org/10.1109/ICRA.2017.7989155

9. Appendices

9.1 Appendix 1: Literature review process

This chapter summarizes the process regarding the literature review that resulted in the literature review of chapter 2. It concerns structured literature research on Scopus and Google Scholar, and the backward snowballing technique based on the papers that were included here.

9.1.1 Structured literature research

The first step for the systematic literature review is identifying a start set of relevant research papers. The choice fell on the combination of two search engines for the search: Scopus and Google Scholar. Scopus.com is owned by the Elsevier and provides an extensive database of peer-reviewed literature in a wide range of scientific and technical disciplines for eligible organizations or individuals (by payment). In addition, Google Scholar is an open access engine that provides an even wider range of potential results, however the results are less strict (Scopus, or more precise the partners of Scopus, use strict rules for the acceptance of papers), which causes that one should be more careful with the acceptance to include papers found by Google Scholar. Another important statement to be made is that papers solely published on arXiv are not excluded directly, although they are not peer-reviewed (yet) when first published here. Nevertheless, arXiv is a widely queried database within the scientific artificial intelligence community, due to the speed that papers are available and the speed that research within this area is evolving.

Scopus

A search query for the starter set has been conducted on *Scopus.com* with the search string “explicability” which resulted in a list of 102 results. The results can be categorized over a variety of subject areas, of which just a part is located within the computer science and engineering category on Scopus. Furthermore, other subject areas could be informative in defining the concept of explicability, so the papers will be analyzed on the added value for this aspect, and if the paper has no added value for this, it will be removed from the list. By selecting the computer science and engineering area, while this is the application area of explicability in the scope of this paper, the list decreases to 43 papers. The list is sorted based on citations from high to low, after which the papers with at least 2 or more citations are included. Applying these scoping criteria, a list of 7 papers remains.

In-text citation:	**Subject:**	**Citations:**	**#** and **backward snowballing**
(Papageorgiou et al., 2006)	Transparency, interpretability, and explicability in tumor grading	98	[1] and no
(T. Chakraborti et al., 2017)	Explicability and model reconciliation problem; the interaction between humans and AI systems	58	[2] and [13]
(Zhang et al., 2017)	Robot planning explicability	39	[3] and no
(Sengupta et al., 2017)	Decision support and explanations	17	[4] and no
(Williams et al., 2018)	Explicability in human-robot interaction	14	[5] and no
(S. Sreedharan et al., 2018)	Explanations of robots and model reconciliation	13	[6] and [14/15/16]
(Adadi & Berrada, 2018)	Black-box models	2	[7] and no

Google Scholar

In addition to the investigation of Scopus, a Google Scholar search on “explicability” resulted in 3800 hits. To scope the search more the search term is made more specific to “explicability” AND “machine learning” which resulted in 2080 results. The first 100 results are investigated, due to the assumption that Google Scholar suggests the most relatable references to these search terms and the ones that have the most citations. This results in the addition of 5 papers to the list.

In-text citation:	**Subject:**	**Citations:**	**#** and **backward snowballing**
(Kannetzky, 2002)	Explicability of speech	19	[8] and no
(Tathagata Chakraborti et al., 2018)	Explicability in task planning	5	[9] and [17/18]
(Sarath Sreedharan, Chakraborti, Muise, & Kambhampati, 2019)	Plan Explicability and Explanations	1	[10] and [19]

(Sarath Sreedharan et al., 2017)	Balancing Explicability and Explanations	7	[11] and no
(Frosst & Hinton, 2017)	Distilling a Neural Network Into a Soft Decision Tree	67	[12] and no

9.1.2 Backward snowballing

With the start-set of 12 papers, a backward snowballing approach has been followed in order to find relevant papers related to this start-set. Backward snowballing is using the reference lists of the start-set to find papers to include (Wohlin, 2014). First, the title is viewed, then the place of the reference and finally the abstract. If the paper still seems useful, you proceed to the full analysis of the paper. When finished with this process for the start-set, it is repeated with the newly added papers until the end of the list is reached. It resulted in the addition of 11 papers.

In-text citation:	**Subject:**	**Citations:**	**#** and **backward snowballing:**
(Kulkarni et al., 2019)	Explicable Planning as Minimizing Distance from Expected Behavior	15	[13] and no
(Miller, 2019)	Explanation in Artificial Intelligence: Insights from the Social Sciences	143	[14] and no
(Lombrozo, 2006)	The structure and function of explanations	355	[15] and no
(Lombrozo, 2012)	Explanation and Abductive Inference	174	[16] and no
(Borgo et al., 2018)	Towards Providing Explanations for AI Planner Decisions	4	[17] and [26]
(Fox et al., 2017)	Explainable planning	59	[18] and no
(Gunning, 2017)	Explainable AI	189	[19] and [20]
(Seegebarth et al., 2012)	Plan explanation framework	48	[20] and no

(Kulesza et al., 2015)	Explanatory debugging, interactive ML	146	[21] and [22]
(Kulesza et al., 2013)	How the soundness and completeness of the explanations impact the fidelity of end-users' mental models	64	[22] and [23]
(Lim & Dey, 2009)	Intelligibility in Context-Aware Applications	115	[23] and no

9.2 Appendix 2: Conducted interviews

Within this thesis process three interviews have been conducted. All the interviews have been recorded and based on this recording transcribed. The transcription method that is used here is the ***edited transcription approach (Streefkerk, 2019); a more concise edited version of an intelligent verbatim transcript***. The advantages of this are that irrelevant sentences, 'fillers' and broken sentences are not included when the meaning of the interview is not changed by it. Furthermore, it is a more formal representation of the interview and it keeps clarity over the content. All the interviews have been anonymized; the interviewee(s) as well as the company. This has been done to protect privacy. The names and companies are known by the interviewer. Due to the fact that the interviews were conducted at different times in the thesis process, the objectives or scope could slightly differ. If an interview was not finished in the time that was scheduled, written answers to the questions were sent afterwards. Editor notes (if needed in the interview) are added between brackets in the transcription: [… …] For future research, more validity could be reached by conducting more interviews and empirical research on the actual use of the framework.

Before the actual start of the interview an introductory conversation on my study, thesis, observations, research goal and myself has been taken place to provide in the creation of mutual understanding on these topics.

Some initial information:

Interview	Appendix 2.1
Interviewer	Nils Herber (NH)
Interviewee(s)	Analytics and Risk Modelling expert (ARM) and an AI Ethics (AIE) expert at a large Dutch retail bank
Date	11-07-2019 15:00
Goal	Confirmatory with respect to the objectives, and exploratory semi-structured interview

Interview	Appendix 2.2
Interviewer	Nils Herber (NH)
Interviewee(s)	Risk Validation (RV) expert and a Model Validation (MV) expert at a large Dutch retail bank
Date	13-08-2019 10:00
Goal	Confirmatory with respect to the objectives, and exploratory semi-structured interview

Interview	Appendix 2.3
Interviewer	Nils Herber (NH)
Interviewee(s)	Business Intelligence (BI) expert at a large supervisory organization for Dutch financial institutions
Date	02-08-2019 15:45

Goal	Confirmatory with respect to the objectives & des. req., and exploratory semi-structured interview

9.2.1 Appendix 2.1 Interview with an Analytics and Risk Modelling expert and an AI Ethics expert at a large Dutch retail bank

NH: Wat doe jij binnen de bank?
ARM: Ik werk bij de afdeling modelling en advanced analytics. Wij zijn gepositioneerd binnen de risk afdeling van de bank. Dan moet je denken aan credit approval modellen, modellen voor balans-sturing, prijsstelling, dat soort vraagstukken hoort al van oudsher bij deze afdeling thuis. Mijn core werkzaamheden houden zich bezig met kwantitatieve modellen bouwen. In principe ben ik bezig in het hele bedrijf met deze modellen dus ook buiten de risk afdeling, bijvoorbeeld modellen waar de output ook gebruikt wordt voor klant interactie. Ik heb veel te maken met GDPR, maar wij willen zelf meer doen dan dat door middel van een data-visie. Ik heb econometrie en wiskunde gestudeerd en ben gepromoveerd op het gebied van statistiek en theoretische econometrie.
De link met jouw afstudeerscriptie is tweeledig. Vanwege mijn werk ben ik een paar jaar terug geïnteresseerd geraakt in de ethiek van data-science, omdat er vaak vragen aan mij gesteld werden hierover. Wij hebben er veelal geen beleid voor op het moment. Banken hebben van oudsher een model governance. Maar dit is niet voldoende om ethische vraagstukken af te handelen. Veel modellen vallen buiten de scope omdat deze volgens de bank niet materieel genoeg zijn, financieel gezien. Deze governance is ook vooral gericht op technische correctheid, terwijl je dingen kan bouwen die technisch correct zijn maar fout voor het gebruik. Ten tweede, er is aan de TU Delft een grant rondom dit onderwerp waar ons bedrijf in participeert.
NH: Goed om te horen dat jullie verder kijken dan de GDPR, dit bespreek ik ook in mijn thesis dat dit nodig is. Daarnaast ben ik ook blij te zien dat je het verschil maakt tussen technische correctheid en ethische correctheid. Ik bespreek dit aan de hand van 3 levels in mijn thesis: machine-level (een model kan technisch goed in elkaar zitten), business-level (kan je de modellen verantwoorden naar de business) en consumer-level (kan je dit uitleggen aan een klant, waaronder een leek). Ik vroeg mij af, zien jullie binnen het bedrijf ook een tweesplitsing tussen uilegbaarheid van modellen tijdens het ontwerpen en achteraf?
ARM: Toevallig heb ik voor mijn vervolgafspraak iets voorbereid. Je hebt de data-scientists die een vraag krijgen van een model-owner (vanuit de business). Op dit moment gaan deze data-scientists dan opschrijven wat ze precies gaan doen en moeten dan een akkoord ophalen bij de model governance committee, bijvoorbeeld voor een credit approval model. Het kan hierbij zijn dat er specifiek op ethische vraagstukken nu nog niet ingezoomd is, omdat dat nog niet in de governance zit. De governance schrijft voor dat na productie dit frequent gemonitord wordt. Hebben we het over andere modellen dan de model governance modellen, dan hebben we daar geen proces voor en hier zijn we ook mee bezig om dit te ontwikkelen. Vooral ook t.o.v. het beschermen van klantbelangen in elk stadium van het proces. Bij het specificeren van de doelstellingen van het model moeten er vragen hierover gesteld worden, en dat de keuzes hier gedocumenteerd worden. Vandaar de Design for Values approach en zijn wij geïnteresseerd ook in dit onderzoek. We willen guidance en een consistent raamwerk creëren voor de afwegingen die hierbij komen kijken.
NH: Bedankt voor deze nuttige uitleg. Op dit moment als er een lening wordt afgewezen, wordt er dan al een uitleg gegeven?
ARM: Op dit moment gaat een klant eerst door bepaalde acceptatiekaders heen. De data komt het systeem binnen en hier komt feedback op. Dit kan een hypotheekadviseur [het voorbeeld hier] inzien en communiceren aan een klant. Bij een afwijzing krijgt de klant alleen te horen dat het model “nee zegt”. Wij vinden dat er bij alle modellen waarbij er een beslissing over of tegen een klant wordt genomen, een uitleg gegeven kan worden (dit staat ook in onze visie. Wij willen transparantie creëren naar de klant toe over wat we doen met de modellen.

AIE: [*komt binnen*] Hallo, sorry dat ik wat later ben.

NH: Goed dat je er bent.
AIE: Laat ik mijzelf even kort introduceren. Ik ben hier aangesteld als AI & ethiek specialist. Ik doe hiernaast ook onderzoek aan een universiteit naar hoe je ethische principes operationaliseert in een bedrijfscontext, in de bankensector. Wij hebben hierbij een aantal onderwerpen waarvan wij zeggen die moeten sowieso aan bod komen, zoals toezicht, AI visie en value-based design.

NH: Laten we naar de objectives gaan voor mijn framework. Deze wil ik graag langslopen en bekijken hoe jullie erover denken en of dit in lijn is met wat jullie denken, en of jullie het nut hiervan inzien. Het framework moet ervoor zorgen dat explicability van beslissingen kan worden verzekerd door middel van een assessment. Ik heb 3 soorten explanations geformuleerd: global ex-ante, global ex-post en local ex-post. [*uitleg wordt gegeven*]
AIE: zit hier nog een situationeel aspect aan vast? Is een bepaalde uitleg nuttiger voor een bepaalde stakeholder?
NH: alle 3 de explanations zouden theoretisch gegeven kunnen worden aan verschillende actoren, maar de ene is nuttiger dan de ander. Laten we doorgaan en gestructureerd de objectives afgaan met de volgende vraag: is de objective nuttig en bruikbaar om toe te voegen?

1. **Enable assessment of … explanations**
    a. **Global ex-ante**

ARM: moet je hebben voor interne purposes. Het is de vraag of het altijd even nuttig is om dit te delen met de klant

    b. **Global ex-post**

ARM: Nuttig voor interne purposes, voor periodieke validie en monitoring

    c. **Local ex-post**

ARM: Nuttig in bepaalde situaties, bijvoorbeeld als een klant verminderde financiele weerbaarheid gaat hebben. Ook voor credit approval, al dan niet met een human-in-the-loop

2. **Should incorporate the situation of unwanted consequences of the decision-making process**

ARM: Ja, bij bijvoorbeeld een lening afwijzen (dit heeft een zware impact) moet er nog een mens naar een uitleg kijken

NH: nu door naar de aspecten van een uitleg. Ik heb er 4 gedefinieerd: completeness, conciseness, soundness and comprehensibility. Hier heb ik ook vier objectives voor, voor het framework.

3. **Ensure completeness of an explanation of a decision**

AIE: een heel nuttige, maar meer completeness leidt niet altijd tot meer vertrouwen bij de gebruiker of de medewerker van het systeem; overcompleteness werkt zelfs tegen vertrouwen. Het moet zo volledig mogelijk zijn voor degene die de uitleg nodig heeft.
ARM: vooral belangrijk voor global ex-ante voor interne doeleinden. Maar heel situationeel.

4. **Ensure conciseness of an explanation of a decision**

AIE: Ja, hier zie ik een spanning, met completeness vs conciseness. Er zal ergens een optimum zitten waarnaar gestreefd moet worden.

5. **Ensure soundness of an explanation of a decision**

AIE: dat is de meest belangrijke. Er moet naar gestreefd worden, maar het kan heel moeilijk zijn om dit te bereiken. Soms kan het in positieve zin beter zijn om iets uit te leggen met een sausje, maar dit moet natuurlijk altijd binnen de wet.
ARM: Alles wat we binnen dit bedrijf doen, hoop ik dat het voldoet aan soundness.

6. **Ensure comprehensibility of an explanation of a decision**

ARM: altijd goed, maar moeilijk om aan te voldoen in deze context.
NH: hier komt altijd een afweging bij kijken tot op welke hoogte je het uitlegbaar moet maken
ARM: het is handig om naar het juridische perspectief hierop te kijken.

7. **Include an iterative component with human explainee to ensure feedback on the given explanations and the ability to adjust the explanation to make it more understandable**

ARM: Dat is nuttig, wij gaan beginnen met het ontwikkelen van een iteratie dashboard voor feedback.
AIE: 1 ding die je wel moet overwegen is hoeveel power je hen moet geven in relatie tot de beslissing en het model. Het mag niet zo zijn dat 3 respondenten het hele model kunnen veranderen. Binnen de crisp-dm moet je niet een te kleine groep de overhand laten nemen over het model

8. **Enable assessment of the cognitive process of producing an explanation**

AIE: Kan handig zijn, maar niet voor klanten speciaal want die mogen niet de overhand krijgen over de programmeurs hierbij.
ARM: heel situationeel, waar ligt de macht, hoeveel 'klanten' zijn er nodig voordat er echt een verandering gaande is. En wie weet hier genoeg van af.

9. **Enable assessment of the social process of transferring an explanation to an explainee**

AIE: vanuit de filosofie gezien is er een discussie tussen mental model groep en pragmatisch omgaan met het model en de beslissing van het model. Ik neig zelf meer naar het praktische deel voor dit sociale proces, en dan kan het handig zijn.
AIE & ARM: hier is ook weer een verschil tussen leren en uitleg rondom een beslissing.

10. **Should be pragmatic to use for decision-makers within the development cycle**

AIE & ARM: Ja, erg belangrijk voor de implementatie van principles

NH: Okee vanwege de tijd en vervolgafspraken van jullie laten we her hierbij dan. Enorm bedankt! Ik zal de overige vragen via de mail opsturen en jullie vragen om deze te beantwoorden.

[Via een email van ARM]

11. **Should be able to be used for prospectively assessment [before the deployment of the system]**

JA, ontwikkelen van een framework wat helpt bij design van algoritme/systeem

12. **Should be able to be used for retrospectively assessment [after the deployment of the system]**

JA, binnen onderneming zelf nuttig om bestaande modellen te toetsen, maar 11 heeft dan wel meer prioriteit; voor auditors gaat 12 natuurlijk helemaal op

13. **Should be able to be used to assess which tasks and decision-making functionalities should not be delegated to the ML system**

JA, voor zoverre dit mogelijk is

14. **Should be delegated to and align with the context variables of the credit underwriting case**

JA, anders zou het raamwerk te abstract blijven

15. **Should include a participatory mechanism to ensure alignment with societal values and understanding of public opinion (iterative feedback with explainee**

JA, maar zou bijvoorbeeld ook kunnen door achteraf verantwoording af te leggen/inzicht te geven

9.2.2 Appendix 2.2 Interview with a Risk Validation expert and a Model Validation expert at a large Dutch retail bank

NH: Hi, thank you for your time and willingness to have this interview.
RV: Yes, companies are very interested in these kinds of topics nowadays.
MV: Thank you, a more general question, is this about answering questions or more a discussion on certain topics
NH: I will first introduce my thesis and then we will move to the objectives and validate this one by one.
MV: What kind of system do you have in mind? One specific type of machine learning system?
NH: The idea is that this framework is agnostic of the machine learning system. It focuses on the explanation of a decision that is made by a machine learning system, and regardless of this ML system and the technique that produces the explanation, the framework should be able to be used to assess the explanation.
RV: So, if I'm correct you see it like this: the bank uses some black-box algorithm that comes up with a decision and you want an explanation that says, this is the decision because of this, this and that.
NH: yes, exactly.
MV: So it is a qualitative assessment?
NH: yes, correct. It should be a qualitative assessment framework to be used by designers of the ML system. With that idea I have formulated objectives for my framework. And I want to discuss them one by one, looking at the relevance, importance, formulation of the objective etc., your opinion on it. Are there any comments on it? My goal is to validate the objectives that I have with your expertise.

A. **Provides guidance for the decision-makers on using the framework**

RV: This comes after the building of the framework I would guess?
NH: yes, correct.
MV: Simplicity and clarity is important here. We are not specifically decision-makers, so I can't tell it for them, but I can tell for myself. If I was, I want something very clear, something quantitative that comes up. Some number or percentage. Ask yourself what is the decision-maker looking for to make this decision.
MV: Make the framework as such that the framework provides the answers that the decision-maker wants.

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

RV: This question is for decision-makers from the business, right?
NH: My idea of VSD theory that I'm using is that designers should think of this already. I know that it works right know that you have the designers and the model validators and then you have the business. However, many researchers in VSD think that already in the design phase these are questions that designers should think about already and I agree with this.
MV: For me, I'm missing the customer within the cycle. The input of them is needed to know if things need to change. And they know the best what explanation suits them the best.
NH: I agree, this is very important, but since this is part of the social process of explanation, and I am focusing on the product of explanation, this aspect is currently out of scope and needs to be investigated in future research. This is a later aspect of the explicability process that definitely needs attention in research.

C. **Is able to prospectively assess explanations**

NH: the assessment needs to take place before the system is deployed, since the problems occur after deployment if there are problems regarding explicability.
MV: This is clear to me.
RV: Yes, same for me.

D. **Is able to assess justificatory explanations**

NH: this is about the format that the explanation takes. With the justificatory value I mean that a decision needs justification, and the explanation is a means to do that. I will definitely include a description of the important words in my framework.
RV: do you think that on average people ask for these features?
NH: I think that it is not the precise point here. If a person asks, the ML system needs to be able to provide a sufficient explanation here.
RV: alright.
NH: do you understand why I focus on the assessment of this justificatory value opposed to a teaching (the explanation on the system) value?
MV: yes, I do.
RV: I agree.

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

NH: this is more related to the linguistic part of an explanation. We should be able to assess the knowledge level and check whether this explanation can be understood.
RV: Yeah
MV: sure, definitely.

F. **Is able to assess the completeness of explanations**

RV: If the available information justifies whether you've got it or not, yes.
MV: The definitions sound overlapping interrelating though.
NH: I will add a firm description of the meanings of the words to the framework, so that the differences are clear. I got these characteristics for quality from this paper [*shows papers*]

G. **Is able to assess the soundness of explanations**

RV: Truthfulness is indeed important to include. It should not be possible that lies are included.
MV: Yes true.

H. **Is able to assess the comprehensibility of explanations**

MV: To me a sound includes comprehensibility, so yes
RV: Yes, to me it seems that they overlap a bit.
NH: I'll make sure that it is clear in the framework how I distinct the definitions.

I. **Is able to assess the conciseness of explanations**

MV: Yes, but this is relative to the required completeness. A layperson needs a concise explanation to be able to keep attention.
RV: I agree.

NH: Thank you very much for all your input! I really appreciate it.

9.2.3 Appendix 2.3 Interview with a Business Intelligence expert at a large supervisory organization for Dutch financial institutions

BI: Verkondigde meningen zijn van mij en niet noodzakelijk van mijn organisatie.
NH: Uiteraard, goed dat je het zegt en geef vooral aan als je iets niet zou willen beantwoorden ook. Hoe zie jij uitlegbaarheid van machine learning?
BI: Uitlegbaarheid zie ik op 2 punten. Uitlegbaarheid van een beslissing over een lening en uitlegbaarheid van het model als geheel.

NH: Die scheiding heb ik inderdaad ook gemaakt. Ik zal me focussen op uitlegbaarheid van uitlegbaarheid van een beslissing omdat ik moet scopen. Je hebt heel veel vormen van uitlegbaarheid
BI: Okee, en op welk moment is het assessment en naar welk moment kijk je?
NH: Het assessment is voor de operationalisatie van het model en kijkt naar een beslissing na de operationalisatie. Op dat moment is deze beslissing en daarbij ook de explanation nog hypothetisch. Dit heet een local ex-post explanation.
BI: Duidelijk. Weet hierbij wel dat het gaat over de operationalisatie van versie 1 van het model. Het blijft een continu proces. Dit wordt continu geëvalueerd. [dit aspect is toegevoegd aan mijn thesis]

NH: Laten we de objectives aflopen met de design requirements die hieruit voortvloeien en de means die het framework zou moeten vormen om aan deze design requirements te voldoen. Deze wil ik valideren. Hiervoor dus voor elke objective en design requirements de vraag: is dit relevant om toe te voegen voor het ontwerpen van het framework? Wat zijn jouw ideeën hierbij?

Objective A: Provides guidance for the users of the framework
Design requirement 1: shall show which new tasks to perform in the development lifecycle
Design requirement 2: shall show how the new tasks need to be performed in the development lifecycle
Design requirement 3: shall show in what sequence the tasks need to be performed in the development lifecycle
Design requirement 4: shall show at what times in the development lifecycle the tasks need to be performed
Means:

- *Includes a user's guideline that explains which new tasks, how, in what sequence and at what times they need to be performed in the development lifecycle.*

BI: Ja, met een framework is het belangrijk om te weten wat te doen, alleen zal een data-scientist er snel mee aan de haal gaan. Hij zal dit veel op eigen inzicht willen doen. Maar een eerste richtlijn is zeker belangrijk. Wil je een mening over wanneer dit framework gebruikt zou moeten worden?
NH: Ja ik maak gebruik van de CRISP-DM cycle en zie dit in gebruik in de evaluatiefase.
BI: snap ik, zo zie ik het zelf ook. Uiteindelijk maak je op basis van het assessment een keuze.

Objective B: Helps with decisions whether a certain task or decision-making functionality can be delegated to the machine learning system
Design requirement 1: shall result in a conclusion if the explanation of interest is good enough
Design requirement 2: shall give an overview of the specific explicability issue(s) in the system if it is not explicable enough
Means:

- *Includes a method to synthesize the assessment results in order to conclude with a decision whether the explanation is good enough, or that it needs to be readdressed, reconsidered and reformulated*
    - *Iterative process, but can't go on forever (sometimes the conclusion is that it is not explicable enough, and another ML system needs to be considered and the data/business understanding phase need to be performed again → crisp-dm)*

BI: Logisch dat je dit wil. Ik zie hierbij wel een punt dat het voor de business veel interessanter is om veel explanations tegelijk te kunnen assessen. [dit heb ik toegevoegd als punt voor future research]

Objective C: Is able to prospectively assess explanations

Design requirement 1: shall be usable for explanation assessment before the system is deployed
Design requirement 2: shall be technique-agnostic, thus usable to assess different formats of explanations
Means:

- *Explicability assessment tasks are included in the evaluation phase of CRISP-DM*
    - *The model(s), in combination with a chosen XAI-technique that fits the model and provides in the required explanation, is (are) scored on explicability, after which the performance is assessed and all the alternatives are compared with regards to the objectives.*
    - *Model and XAI technique combination with sufficient explicability, and preferably the highest, and the best performance is chosen. If this combination does not reach the threshold(s), one should iterate back to business understanding*
- *Is not model-specific; i.e. can assess the explanations on model-independent criteria (include this)*

BI: Dus het framework is onafhankelijk van de methode voor machine learning en de methode om een explanation te formuleren. Ja lijkt me verstandig. Een kanttekening hierbij is dat DQ1 vooral theoretisch is. Voordat het systeem deployed is, is tegelijkertijd erna, doordat het continu verbeterd wordt na deployment.
NH: Dus het is net hoe je deployed definieert.
BI: Ja precies, wees duidelijk dat je het over een versie 1 heeft. Er is niet iets als een 'final versie'.

Objective D: Is able to assess local explanations
Design requirement: shall be usable for assessment of explanations on decisions
Means:

- *Assessment criteria are specified towards explanations on decision instead of the system*
    - *Not all aspects of the system (completeness)*

BI: Dit volgt direct uit je scoping als ik het goed heb.
NH: Klopt ja.
BI: Dit is inherent uit je scope.
NH: Waarom ik hem toch heb toegevoegd op dit moment is dat we later de completeness bespreken. Hier zijn allerlei onderdelen die toegevoegd kunnen worden, en deze keuze bepaalt welke aspecten bij completeness erin moeten zitten.

Objective E: Is able to assess ex-post explanations
Design requirement: shall be usable for explanations after a (hypothetical) decision has been made
Means:

- *Assessment criteria are specified towards explanations after the moment that a decision has been made*

BI: Volgt uit local, dus voeg dit samen met objective D. [dit is gedaan] Het hoeft voor mij geen eis te zijn voor je framework aangezien het direct uit je doelstelling komt [deze objective en objective D zijn hierom later eruit gehaald]

Objective F: Is able to assess justificatory explanations (in order to enhance accountability)

Design requirement 1: shall be able to check which 'evidence roles' are useful in the explanation
Design requirement 2: shall be able to check which 'evidence roles' are present in the explanation
Means:
Includes the question for every evidence roles (Biran & McKeown, 2014) that are required to be included:

- *Does the explanation contain *X- evidence role*?*

BI: Justification is belangrijk, maar ik kan dit requirement niet assessen omdat het voor mij nu niet duidelijk is wat de evidence-roles precies betekenen.

Objective G: Is able to assess consumer-level explanations
Design requirement: shall be able to specify the knowledge level or expertise that the explainee needs to have in order to understand the explanation
Means:

- *Includes the question: What knowledge does the explainee need to have in order to understand the explanation?*

BI: Hoe beantwoord je deze vraag?
NH: Als er een hypothetische beslissing is met een uitleg, zou je met deze vraag kunnen bekijken of er een te technisch begrip in zit die wellicht simpeler opgeschreven kan worden.
BI: Is handig om naar te kijken. Maar dit is vooral een aspect wat naar voren moet komen in het sociale proces. De vraag stellen is een goede, maar dit moet empirisch gevalideerd worden ook. Ik zou dit nooit zelf inschatten
NH: Goed punt, dit heb ik zelf ook al genoemd in.

Objective H: Is able to assess explanations towards a layperson
Design requirement: shall be able to decide if an explanation is understandable enough for a layperson
Means:

- *Includes the question: Is the explanation understandable for a person that has no expertise in- or knowledge of credit underwriting, personal loans, machine learning nor automated decision-making?*

BI: Wat is het onderscheid tussen H en G?
NH: Het onderscheid is dat een layperson alleen baat heeft bij een consumer-level explanation, maar dat een consumer-level explanation nuttig kan zijn voor alle vormen van explainees. Zo kan een auditor ook een consumer-level explanation nodig hebben om deze te contoleren.

BI: Ik zou ze samenvoegen aangezien het antwoord op vraag H het antwoord op vaag G geeft. Goed punt dat je de open vraag stelt wel, en het is wel een belangrijk punt. [deze 2 zijn samengevoegd]

Objective I: Is able to assess the completeness of explanations
Design requirement 1: shall be usable to check which intelligibility types the explanation requires to fulfill a sufficient level of completeness
Design requirement 2: shall be usable to assess the extent to which the relevant aspects are included in the explanation
Means:

- *Includes a conclusion on what intelligibility types are relevant for the type of explanation to be assessed; does the explanation need to incorporate this type?*

(Lim & Dey, 2009)

- *Application - Input: If there is high externality (Dependency on external sources) users are more interested in the input sources and readings*
- *Application - Output: the output alternatives, may be of interest to investigate multiple recommendations*
- *Application – Model - Why: informative on the reason for a decision that satisfy the users' inquiries*
- *Application – Model - How: users may like to know how the application arrives at its outcomes*
- *Application – Model - Why not: risk of inappropriateness, goal-supportive functions; why not another possibility*
- *Application – Model - What if: what if this is changed, has this an outcome on the decision?*
- *Application – Model - What else: when users are aware of a certain availability, this could be of interest*
- *Application – Model - Visualization: could be interesting to provide, to improve effectiveness and comply to demand*
- *Application – Model - Certainty: goal-supportive applications, how certain is the decision (probabilities).*
- *Application – Model - Control: would support users changing parameters in the model, in order to investigate the application*
- *Situation: increases the real-world situational awareness; e.g. the historical trace of events, related events and contexts. More critical in highly critical situations*

- *Includes the following question on all these types that are considered as relevant:*
  - *Is this intelligibility type addressed in the explanation?*

BI: Hier moet inderdaad een selectie op gemaakt worden, want er zal geen uitleg nodig zijn die alle types als onderdeel heeft. Een local ex-post explanation heeft niet alle types nodig zijn. Als je een selectie maakt is dit een handige manier om dit zo gestructureerd af te gaan.

Objective J: Is able to assess the soundness of explanations
Design requirement: 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
Means:

- *Is the explanation a correct representation of how the model came to the decision?*
  - *Is this explanation aspect based on the full truth, a simplified truth model or the truth of a singular feature, or not the truth?*

BI: Jouw framework moet dus de data-scientist inzicht geven dat de uitleg gebaseerd is op de waarheid?
NH: Op het moment dat de machine-level is gevalideerd en de business-level soundness is gevalideerd kan je de overeenkomst van de explanation met deze waarheden vergelijken. Je kijkt dus niet naar de validiteit van het model met deze vraag maar naar de validiteit van de uitleg ten opzichte van het model.
BI: Dus aangenomen dat de machine-level correct is en de business-level correct, kijk je naar of de uitleg een correcte vertegenwoordiging is van hoe het model tot de beslissing gekomen is? Ja dit is een belangrijke toevoeging, want het moet op de waarheid gebaseerd zijn.

Objective K: Is able to assess the comprehensibility of explanations
Design requirement 1: shall be usable for the assessment of the narrative aspect of an explanation
Design requirement 2: shall be usable for the assessment of the human-interpretable linguistic level of the explanation
Means:

- *Includes the question:*
    - *Does the explanation include singular facts or datapoints without context around it?*
    - *Is the explanation textually written in a narrative format?*
    - *Is the language used considered easily understandable for humans?*
    - *Does the explanation sufficiently link the decision, important features, the roles and effects of these features and important values in a logical way?*

BI: Je kijkt hier alleen naar taal?
NH: Ja klopt, dit zal ik duidelijker maken in mijn thesis. [Tekstdeel dat er alleen naar textuele uitleg in het engels wordt gekeken is toegevoegd aan thesis]
BI: Ja dan is het een duidelijk punt.

Objective L: Is able to assess the conciseness of explanations
Design requirement 1: shall be able to assess the explanation length
Design requirement 2: shall be able to assess the number of concepts included in the explanation
Design requirement 3: shall be able to assess the modularity of the explanation structure
Means:

- *Includes the questions:*
    - *How many textual lines does the explanation include?*
    - *How many words does the explanation include?*
    - *How many concepts are included in the explanation?*
    - *Is the explanation structure considered modular?*
        - *I.e. can the explanation easily be extended when consumers ask for more reasoning and justification, without losing the structure of the explanation?*

NH: dit zal afhankelijk zijn van de situatie en vraag van de layperson, en zal dus meegenomen moeten worden in het sociale proces van uitlegbaarheid.
BI: Ja begrijpelijk, verder wel een goed punt om mee te nemen in je framework.

NH: Heel erg bedankt voor alle informatie! Ik heb er heel veel aan gehad.

9.3 Appendix 3: EAF User's Guidelines

9.3.1 Important statement regarding the adoption of the framework

This framework is designed for the following context characteristics (See 2.3.1.4):

- Textual explanation product (explicability process sub-part)
- Prospective (assessment moment)
- Justification (explanatory value)
- Local (explanation scope)
- Ex-post (moment in time)
- Consumer-level (level of abstraction)
- Layperson (explainee)

My hypothesis is that parts of the framework can be adopted in future frameworks for other (slightly different) context characteristics. Future research should show to what extent parts of this framework can be adopted in a framework for a case with different context characteristics, but until this moment it should be used for use-cases with similar context characteristics.

9.3.2 Introduction

These user's guidelines are meant as an explanation and hands-on advice on how the framework should be used. These guidelines concern the Explicability Assessment Framework (EAF) as designed in this thesis.

The goal of the use of the framework is to *assess the explicability of an explanation of a machine learning system's decision.* The framework enhances the overview of the machine-learning system use case and forms a tool for the structured evaluation of explanations on decisions, in order to create handles for decision-makers in the development lifecycle of the machine learning system. These handles help the decision-makers to decide whether a certain explanation is good enough, which enhances the explicability of a machine learning system.

Chapter 8.3.3.2 shows the integration of the assessment with the CRISP-DM framework (at what time in the cycle does the assessment need to take place), the sub-sections of the framework, what tasks need to be performed in these sections, how the tasks need to be performed and in what sequence.

9.3.3 Guidelines for Explicability Assessment Framework

9.3.3.1 The Explicability Assessment Framework (AEF)

Table 10: Explicability Assessment Framework (AEF)

1. <u>Context characteristics</u>		
Decision type	*insert description of the type of decision that the machine learning system makes in the case of interest*	
Value(s)	*insert the value of interest for the design of the machine learning system in the use case*	
Actor-forum relationship	*insert a description of the type of relationship the actor and the forum have*	
Forum	*insert the forum of the decision*	
Actor	*insert the actor of the decision*	
Potential effects	*insert a description of the potential effects for the forum and actor of the decision*	
Ethical need	*insert a description of the ethical need that drives incorporating the value in the design of the system*	
Regulation(s)	*insert a description of the regulation(s) that drives incorporating the value in the design of the system*	
2. <u>Explanation characteristics</u>		
Explainee	Layperson/business/data-scientist/auditor	
Level of abstraction	Consumer-level/business-level/machine-level	
Explanatory value	Justification/teaching	
Explanation scope	Local/global	
Moment in time	Ex-post/ex-ante	
Explicability process sub-part	product/cognitive process/social process	
3. <u>Framework adjustments</u>		
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	

4. <u>Assessment Object Assessment</u>	
insert the full explanation to be assessed	
Questions:	***Answers:***
Justificatory explanation	
1.X For evidence role X, does the explanation contain this evidence role?	
Explanation towards a layperson	
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?	
Completeness	
3.X For intelligibility type X, does the explanation contain this intelligibility type?	
Soundness	
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?	
Comprehensibility	
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?	
Conciseness	
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 … *	

9.3.3.2 Integration in CRISP-DM lifecycle

The assessment framework needs to be aligned with the development cycle of the company who uses it. This thesis adheres to the widely accepted development lifecycle (in the industry) of CRISP-DM (Wirth & Hipp, 2000). Looking into this, the evaluation phase includes the task of *evaluating results* that give as output *an assessment of the Data Mining Results with respect to Business Success and the Criteria*. Moreover, it provides in the final *approved models.* Since we introduce a new type of assessment regarding the models, i.e. the assessment of formulated explanations on decisions of these models, and the approval of these models is dependent as well on the goodness of these explanations, **the explicability assessment needs to be incorporated in the evaluation phase of the CRISP-DM.**

In addition to the former '*evaluate results'-task* additional tasks are added before the model(s) can be approved. We assume that designers within this development lifecycle will only consider machine learning models that have significant increase in prediction-accuracy over the former less complex models to evaluate explicability. Therefore, the first iteration of approving models has taken place. These models will not move on to the deployment phase, but now have to be assessed on explicability as well. The sufficiently explicable model(s) can move on to the deployment phase.

If no model can be found with a sufficient explicability level and increase in prediction-accuracy, one should iterate back to the business understanding phase (again: aligned with the CRISP-DM cycle). Here, the tasks should be readdressed, reconsidered and reperformed, as well as the other phases until evaluation, and if still no model can be found with a sufficient explicability level and increased prediction-accuracy, the users of the framework should conclude to stick with the old less complex model that has a high explicability level.

9.3.3.3 Context characteristics

Step	Task
1.1	Describe the type of decision that is made by the machine learning system in the use case. I.e. Who or what is the subject of the decision, what is decided on and in what use case?
1.2	Describe what the main value(s) of interest is (are) to be incorporated in the machine learning system in the use case
1.3	Describe what the type of actor-forum relationship is that this value (or these values) implies
1.4	Describe who the forum is of the relationship
1.5	Describe who the actor is of the relationship
1.6	Describe what the potential effects are for the forum and the actor of the decision by the machine learning system
1.7	Describe the ethical need of that drives the urge to incorporate the value(s) in the machine learning system
1.8	Describe the regulation(s) that drives the urge to incorporate the value(s) in the machine learning system

Additional explanation and/or example:

Step	Explanation
1.1	E.g. the machine learning system decides on the acceptance or denial of a credit application for a personal loan in the European Union
1.2	E.g. explicability as a means for accountability
1.3	E.g. public accountability relationship between the bank and the credit applicant (consumer)
1.4	E.g. the forum is the subject of the decision, or the credit applicant (Accountee)
1.5	E.g. the actor is the bank or the decision-maker on the rejection or acceptance of the credit application (accountor)
1.6	E.g. receiving/being denied a loan, having more good outstanding loans/less bad outstanding loans, risk of discrimination and as a consequence: legal prosecution, or publicly released news-item(s) that could cause damage to the brand, or distrust
1.7	E.g. consumers want to know what the reasoning is behind a certain decision that they are subject to

1.8	E.g. GDPR and CCD asks for the possibility to justify and explain a decision on the application of a loan

9.3.3.4 Explanation characteristics

Step	Task
2.1	Decide on who the explainee is of the explanation to be assessed, with regards to the forum of the relationship
2.2	Decide on what the level of abstraction is of the explanation to be assessed
2.3	Decide on the explanatory value of importance for the explainee
2.4	Decide on the explanation scope with regards to the explanatory value of the explanation
2.5	Decide on the moment in time with regards to the explanatory value and the explanation scope
2.6	Decide on the explicability process sub-part to assess

Additional explanation and/or example:

Step	Explanation
2.1	Choose: Layperson, business employee, data-scientist, auditor - Data scientists: the developers, programmers, and modelers of the machine learning system that have sufficient mathematical and technological knowledge in such a way that they can understand a more technical and mathematical explanation of the machine learning system. - Business employees: the employees within a company that have knowledge of the business objectives and requirements of the project. 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. In addition, they are the ones that have contact with the consumer if a consumer wants an explanation concerning the decision-making with a loan application. - Layperson (consumer): the loan applicants that want to know why a certain decision was made. This is a broad group with a wide range of expertise (from no expertise, or

	layperson, to a high level of expertise, or expert). Thus the explanations regarding the decision need to be very accessible and understandable. - Auditors: the external controllers and validators of the machine learning system. They are professionals in the field and have a firm understanding of machine learning systems and data governance, and the problems that can occur with them. They have most interest in the underlying system and whether it is compliant to certain standards and principles (that are arguably not defined currently)
2.2	Choose: machine-level, business-level, consumer-level - Machine-level concerns a mathematically sound explanation so that a data scientist and/or mathematician can validate the quality of the algorithms, calculations and model outcomes; i.e. the bottom abstraction level. Hereby the explainers and the explainees are both humans from the data science domain who are required to have some proficiency with regards to machine-interpretability. - Business-level relates to 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; i.e. the middle abstraction level. 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 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; i.e. the top abstraction level. 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 (consumer).
2.3	Choose: justification, teaching - Justification: the explanation is focused on justifying (to give a good reasoning for) a decision or multiple decisions - Teaching: the explanation is not focused on a decision but on the transfer of knowledge on the workings of (a sub-system of) the machine learning system

2.4	Choose: local, global - Local: explanation of a specific decision - Global: explanation of the system functionality
2.5	Choose: ex-post, ex-ante - Ex-post: after an (hypothetical) automated decision - Ex-ante: prior to an automated decision
2.6	Choose: explanation product, explanation cognitive process, explanation social process - explanation 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 product: the explanation that results from this cognitive process is the product (e.g. a textual, visual or conversational explanation) - explanation 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

9.3.3.5 Framework adjustments

Evidence roles selection

Step	Task
3.1	Decide on which evidence roles need to be included in a good explanation within the scope of the explanation and context characteristics

Additional explanation and/or example:

Step	Explanation
3.1	To assess the justificatory aspects of an explanation, we make use of the evidence roles, as defined by Biran & McKeown (2014): - 'normal evidence' o E.g. evidence expected to be present in many instances predicted to be in this class (high positive importance, high positive effect on the prediction) - 'exceptional evidence'

    - E.g. evidence that is not usually expected to be present (low importance, high positive effect on the prediction)
- ‘contrarian evidence’
    - E.g. strongly unexpected evidence, since the effect has the opposite sign than expected (high negative importance, high positive effect on the prediction)
- ‘missing evidence’?
    - E.g. important features that were expected to contribute highly positively on the prediction, but were weak for the prediction (high positive importance, low effect on the prediction)
- ‘normal counter-evidence’
    - E.g. expected to contribute negatively on the prediction and does contribute negatively on the prediction (high negative importance, high negative effect on the prediction)
- ‘exceptional counter-evidence’
    - E.g. unexpected, the feature is not expected to contribute highly negative on the prediction, but it does contribute highly negative (low importance, high negative effect on the prediction)
- ‘contrarian counter-evidence’
    - E.g. feature we expect to contribute positively, but contributes negatively instead (high positive importance, high negative effect on the prediction)
- ‘missing counter-evidence’
    - E.g. feature that was expected to contribute highly negatively, but was weak for the prediction (high negative importance, low effect on the prediction)

Intelligibility types selection

Step	Task
3.2	Decide on which intelligibility types need to be included in a good explanation within the scope of the explanation and context characteristics

Additional explanation:

Step	**Explanation**

3.2	To assess completeness, Lim & Dey (2009) provide a comprehensive list of different types of information (intelligibility types) that will be used for the assessment, that could be included in an explanation, and this list is used in the investigation of Kulesza et al. (2013) as well: • **Input**: If there is high externality (dependency on external sources) users are more interested in the input sources and readings. Information on the input of the model. • **Output**: the output alternatives; may be of interest to investigate multiple recommendations • **Why**: information on the reasons for a decision that satisfy the explainees' inquiries • **How**: explainees may like to know how the application arrives at its outcomes • **Why not**: risk of inappropriateness, goal-supportive functions; why not another possibility? • **What if:** what if X is changed, has this an effect on the decision? • **What else**: when users are aware of a certain availability, this could be of interest; is there any other available option? • **Visualization**: could be interesting to provide, to improve effectiveness, clarity and comply to demand of the explainee • **Certainty**: goal-supportive applications, how certain is the decision (probabilities)? • **Control:** would support users changing parameters in the model, in order to investigate the application. • **Situation**: increases the real-world situational awareness; e.g. the historical trace of events, related events, and contexts. More critical in highly critical situations?

9.3.3.6 The assessment

Step	Task
4.1	Include the formulated explanation that will be assessed in the framework
4.2	For every evidence role defined in 2.3.1, answer the following question: does the explanation from 2.4.1 contain this evidence role?
4.3	Answer the following question: what knowledge does the explainee need to have in order to understand the explanation?

4.4	Answer the following question: Is the answer to 2.4.3 conform the corresponding knowledge of the explainee?
4.5	For every intelligibility type defined in 2.3.2, answer the following question: does the explanation contain this intelligibility type?
4.6	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?
4.7	Answer the following question: Is the explanation textually written in a narrative format?
4.8	Answer the following question: Does the explanation include singular facts or datapoints without context?
4.9	Answer the following question: Is the used language considered easily understandable for humans?
4.10	Answer the following question: Does the explanation sufficiently link the decision, important features, the roles and the effects of these features in a logical way?
4.11	Answer the following question: How many textual lines does the explanation include?
4.12	Answer the following question: How many words does the explanation include?
4.13	Answer the following question: How many concepts are included in the explanation?
4.14	Answer the following 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?

Additional explanation and/or example:

Step	**Explanation**
4.1	Transparency and overview of the assessment will improve if the to-be assessed explanation is documented in the framework as well
4.2	The evidence roles that are relevant to the consumer are valuable for the justificatory goal of the explanation
4.3	An open question like this should give insight into the background that an explainee should have to understand the reasoning in the explanation

4.4	If the background required is not in line with the proposed explainees' background, there is a mismatch and the explanation is considered less understandable, which could cause problems
4.5	The relevant intelligibility types to include improve the completeness of the explanation, if they are included
4.6	The explanation should correctly represent (parts of) the decision-making of the model, and there should be no mistakes in here. The complete soundness of the full system can only be ensured if and only if the soundness of the machine-level and business-level is ensured. Nonetheless, the assessment of the soundness of the consumer-level explanation can ensure the validity of the explanation with respect to the known workings of the system.
4.7	The textual format of an explanation is the scope of this framework and a narrative format increases the understanding of the explanation
4.8	Singular facts or datapoints without context leaves a lot to the imagination of the explainee, which could result in incorrect interpretation
4.9	The use of the most common and basic words to describe a certain concept improves the comprehensibility of an explanation, so a check needs to be in place if other more basic words could be used to improve this, including a check if mathematical operators and non-linguistic signs and language is used (since this is not understandable for everyone).
4.10	The logical linkage of the decision, the important features, their roles and their effects on the decision improves the coherence of the explanation, which improves the comprehensibility of it
4.11	A shorter more concise explanation improves the attention of the explainee towards the explanation, which enhances the understanding of the explanation
4.12	" " " " " " "
4.13	The number of different concepts that are included in the explanation increases the amount of time that the explainee has to think about the explanation to understand it. This is not necessarily directly a bad thing, however fast understanding of an explanation is preferable.
4.14	A modular structure can more easily be extended, and this could enhance the possibility to better serve the requirements of the explainees in an iterative social process

9.3.3.7 Concluding section [5]

Taken all the steps represented in these guidelines, an overview has been formed on the status of explicability of the machine learning system, regarding the specific context and explanation that has been assessed.

The answers of the questions in the assessment should be evaluated on whether these answers show the achievement of a sufficient level of explicability on these assessment criteria, after which this should be documented in section 5 of the AEF (the concluding section). If a sufficient level is not reached, the conclusion should include the conducted steps within the assessment that imply an insufficient level of explicability. The next step is to iterate back towards the business understanding phase and move along the other phases in order to investigate if it is possible to improve the explicability. If it seems that it is not possible, one can conclude that the machine learning system cannot be deployed to take on this decision-making functionality, due to the lack of explicability.