Guiding the specification of sociotechnical Machine Learning systems

Addressing vulnerabilities and challenges in Machine Learning practice

Master thesis

Complex Systems Engineering and Management

Faculty of Technology, Policy and Management

by

Anouk Wolters

Student number: 4476239

Defence: May 10th, 2022

Graduation Committee
Chairperson: Prof.dr.ir. M.F.W.H.A. Janssen, Engineering Systems and Services
First Supervisor: Dr.ir. R.I.J. Dobbe, Engineering Systems and Services
Second Supervisor: Dr. F.S. Gürses, Multi-Actor Systems
External Supervisor: MSc. N. Jetten, Deeploy

Preface

Here it is: my master thesis on guiding the specification of sociotechnical Machine Learning systems! I hope it provides insight in the challenges that the development and operations of Machine Learning systems bring, both from a scientific and an empirical point of view. The thesis aims to take a first step in establishing a sociotechnical view in the specification of these systems, by providing ten guidelines to Machine Learning practice.

Since I switched to the ICT track at the beginning of my CoSEM master, and at the same organised a conference with a theme related to Artificial Intelligence, my interest in the Artificial Intelligence field has been sparked. When Roel connected me with the startup Deeploy for a potential thesis internship, this brought everything together for me. I got to experience working in a young, entrepreneurial setting, while applying my CoSEM view on this interesting field.

To everyone within Deeploy: Thank you for this opportunity and the great time! I felt very welcome from the beginning and being part of the company has been both inspiring and fun! To Nick, thank you for being my supervisor on behalf of Deeploy. You were always happy to help connect me to people I needed to speak to, provide me with feedback, and meet with me every week to discuss my work.

From the TU Delft, I first would like Marijn. Thank you for being the Chair of the committee, providing me with insightful feedback and clear directions on how to proceed after all official meetings. Second, I would like to thank Seda. Your critical but constructive reflections and sharp notes on my work really pushed me to think a step further every time. At the same time, you really helped to determine what is feasible for a master thesis and what is not, which really helped to keep this research manageable. Lastly, I really appreciated that you checked in if I was still doing fine at every meeting. Last but not least, Roel. Your enthusiasm really helped me to go through this process. Writing a thesis is a little lonely sometimes, but your positiveness and input always helped me to continue and become more confident about the work I was doing. You made time for me every week, and even when you were on holiday in the last crucial weeks of writing my thesis, you made time to meet with me online. Our brainstorms and discussions always helped me when I was feeling stuck, or had a lot of ideas I was not sure of. Thank you for everything, Roel!

The last couple of months have definitely been a challenge, filled with ups and downs. All in all, it has been a rewarding experience in which I certainly learned a lot. I hope you will enjoy reading this thesis!

This also marks the end of the great time I have had the last 6,5 years spent in Delft. I look back with very warm feelings to everything I have learned, experiences I have had and friends I have made. The student life is over, but great things will certainly follow.

Lets's see what the future brings!

Anouk Wolters
April 26, 2022

Executive summary

Experts expect that there will be a 'Fourth Industrial Revolution' in the next few decades. The fundamental shift will be in decision-making, as computers are increasingly able to make reliable decisions, whereas before only humans were capable of decision-making. This shift is caused by Artificial Intelligence (AI). Within the range of AI technologies, Machine Learning (ML) models are mathematical models that can learn from examples, to identify relevant patterns in data sets. This way, ML models can make predictions or decisions without being explicitly programmed to perform the task. ML models are increasingly being integrated into decision-making processes, as they have the potential to increase efficiency and improve decision-making. However, ML systems can be in violation with fundamental rights and may lead to physical dangers. Furthermore, their output can result in economic losses to organisations or citizens. These harms imposed by ML systems are primarily characterised as 'bias' or technical flaws in the design of the technical system, which leads to a focus on technical solutions. This way, a complex sociotechnical problem is narrowed down to a problem in the technical design of ML systems, and thus in the hands of technology companies and technical stakeholders. On the other hand, AI ethics initiatives have proposed high-level principles and statements lately, but how to translate these to the specification of ML systems in practice remains unclear.

There is a need for a more comprehensive sociotechnical systems view on ML. Such a view looks at the development and use of an ML system in practice as being a sociotechnical ML system: "a system consisting of technical artefacts, human agents and institutions, in which a machine-based subsystem influences its real or virtual environment by automating, supporting or augmenting decision-making". This research takes on this view to design a sociotechnical guide for ML practice, centring the specification of sociotechnical ML systems. Taking on the guidelines contributes to a safe and effective development and use of ML systems. Before this sociotechnical guide could be designed, two identified knowledge gaps had to be filled. First, a comprehensive overview of vulnerabilities that emerge in sociotechnical ML systems and how to synthesise these in sociotechnical ML system's dimensions, was lacking. The second research gap shows that there was little insight in how is dealt with dimensions of vulnerabilities in the development and use of ML systems in practical settings.

The research objective is to design pragmatic guidelines that can be used as a starting point in the development of ML use cases, that centrers the sociotechnical specification of sociotechnical ML systems. A Design Science Research (DSR) approach is used to answer the main research question to meet the research objective, which is formulated as follows:

What guidelines should be followed in ML practice to establish a sociotechnical ML systems view in the specification of Machine Learning systems?

First, the knowledge base for the research was laid out by conducting an integrative literature review. This resulted in the construction of a theoretical framework, that synthesizes the sociotechnical ML system's dimensions in which vulnerabilities emerge. These dimensions are: *Misspecification*, *Machine error*, *Interpretation*, *Behaviour*, *Adaptation*, *Dynamic change*, *Downstream impact*, and *Accountability*. Next, empirical insights are gathered in a local practice environment. To gather those, semi-structured interviews are conducted with eleven stakeholders involved in two ML use cases within two different banks in the financial sector, as well as seven interviews with representatives of civil society organisations and regulatory bodies. Based on an inductive analysis of these interview data, seven main challenges that occur in ML practice are identified, using a sociotechnical ML system perspective. Those challenges occur along the dimensions of the theoretical framework, experienced in different contexts by different stakeholders. The seven challenges are:

- Challenge 1: Defining the system boundaries for the ML lifecycle
- Challenge 2: Dealing with emergent dimensions in the sociotechnical ML system
- Challenge 3: Introducing a human decision-maker in an ML decision-making process
- Challenge 4: Recognising the importance of data in the ML lifecycle
- Challenge 5: Developing knowledge and communication within sociotechnical ML systems
- Challenge 6: Providing transparency of ML systems and outcomes
- Challenge 7: Interpreting regulations applicable to ML systems

The insights from the knowledge base and the local practice environment ultimately lead to the design of a sociotechnical guide, consisting of ten guidelines with proposed directions for implementation and a schematic sociotechnical ML lifecycle process, as an alternative for the currently used ML lifecycle. The guidelines aim to contribute to solving the seven challenges and are associated with the eight dimensions described above. The following guidelines have been developed:

- Guideline 1: Establish a multidisciplinary team at the beginning of the ML lifecycle
- Guideline 2: Define the system boundaries as multidisciplinary team
- Guideline 3: Enable the identification, addressing, and mitigation of vulnerabilities in the sociotechnical specification
- Guideline 4: Formulate an initial specification of the sociotechnical ML system before starting the experimental stage of the sociotechnical ML lifecycle
- Guideline 5: Create feedback channels for different stakeholders during the development and operations of the sociotechnical ML system
- Guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical ML system in operation
- Guideline 7: Verify and validate the sociotechnical ML system before operationalizing
- Guideline 8: Establish transparency of the sociotechnical ML system
- Guideline 9: Create knowledge and communication between stakeholders in the sociotechnical ML system
- Guideline 10: Establish a safe culture and adequate management within the organisation

The results of the research contribute to ML practitioners, as the guidelines provide a starting point to ML uses cases that establishes a sociotechnical ML systems view. Ultimately, following the guidelines lead to safer and more effective ML development and operations. Further, the results contribute to civil society organisations and regulatory bodies as the theoretical framework supports the articulation of vulnerabilities they see in the ML field, as well as to bridge their concerns to policymakers and ML system developers. Lastly, the research provides insights to the company Deeploy, with whom the researcher collaborated. Guidelines 3, 5, 6, 7, and 8 can (partly) be accommodated by the Deeploy platform. However, caution is needed, because besides the contributions that implementing a technical platform as Deeploy for ML operations brings, the other guidelines should be followed as well to establish safe and effective sociotechnical ML systems. These guidelines require more organisational effort.

The research also contributes to the scientific knowledge base, as the developed theoretical framework can be used in future research on vulnerabilities in sociotechnical ML systems. Moreover, the empirical insights gathered in this research provide directions for research in which actual challenges encountered in practice can be addressed.

Three directions to build upon this research are recommended. First, evaluation and demonstration of using the guidelines in real-life ML use cases is recommended. The design cycle in this research did not involve the required iterative steps between design and evaluation, but concluded with a first design of guidelines. Evaluation and demonstration are recommended to research the value of using the guidelines in actual ML use cases and to iteratively improve the guidelines.

Second, researching other ML use cases within different organisations within the financial sector and in other sectors would enrich the research output. As the financial sector is highly regulated and risk averse, it would be insightful to include use cases from less regulated and more risk seeking sectors, to complement this research. Lastly, this research widened the technical view that dominates the ML field towards a sociotechnical ML system view. The next step is to widen the sociotechnical ML system view even further, by centralizing the larger organisational and societal changes that are caused by the introduction of ML systems in research.

List of Figures

List of Tables

Abbreviations

ADM Automated decision-making
AI Artificial Intelligence
API Application Programmable Interface
CD Continuous Delivery
DDDM Data driven decision-making
CI Continuous Integration
CT Continuous Training
CoSEM Complex Systems Engineering and Management
Dev Development
DNB De Nederlandsche Bank
DSR Design Science Research
DSS Decision support systems
EU European Union
GDPR General Data Protection Regulation
ML Machine Learning
MLOps Machine Learning Operations
Ops Operations

Contents

Chapter 1

Introduction

1.1 Context: Artificial Intelligence and Machine Learning in high stakes domains

Experts expect that in the next few decades, the 'Fourth Industrial Revolution' will take place (Syam & Sharma, 2018). The fundamental shift will be in decision-making, as computers are promised to be able to make reliable decisions, whereas before only humans were capable of decision-making (Syam & Sharma, 2018). The technology accommodating this shift is Artificial Intelligence (AI). The High-Level Expert Group on AI defined AI as: "Systems that display intelligent behaviour by analysing their environment and taking actions - with some degree of autonomy - to achieve specific goals." (AI HLEG, 2019). AI can bring a wide array of economic and societal benefits across all industries and social activities (European Commission, 2021). The technology is rapidly adopted in high stakes social domains, and reshapes many public, professional, and personal domains (Dobbe, Krendl Gilbert, & Mintz, 2021). Whilst AI has the potential to increase efficiency and improve decision-making, it can lead to harms (Balayn & Gürses, 2021). AI systems can be in violation with fundamental rights with regard to discrimination, privacy or stereotypical representations (Balayn & Gürses, 2021). Other harms include physical dangers related to new robotic systems as autonomous vehicles, and welfare systems leading to economic losses (Dobbe et al., 2021).

In this master thesis, the focus will be on Machine Learning (ML). ML is a subfield of AI which "builds a mathematical model based on sample data, known as "training data," in order to make predictions or decisions without being explicitly programmed to perform the task." (Zhang, 2020, p. 223). Instead of being simple rule-based models, ML models learn from examples (Rajkomar, Dean, & Kohane, 2019). ML models identify relevant patterns in data sets, in order to turn them into usable information for decision-making (Veale & Brass, 2019). ML models are increasingly being integrated into important decision-making processes, to improve decision-making (Green & Chen, 2020). Examples are judges using risk assessments to determine criminal sentences, health departments within municipalities using algorithms to prioritise inspections, and banks using models to manage credit risk (Green & Chen, 2019b). This is a major shift in decision-making: where decision-making was a social enterprise originally, it has become a sociotechnical event (Green & Chen, 2019b).

1.2 Introducing ML concepts and a sociotechnical ML systems view

To further define the scope for this master thesis, it is important to arrive at a mutual understanding of some core concepts. These concepts are introduced in the subsequent sections.

1.2.1 Data, Algorithms, ML models, ML systems

An ML algorithm is the set of calculations to perform to arrive at an ML model that will perform inferences about the future (Balayn & Gürses, 2021). An algorithm identifies the main patterns in available data and guides the learning of inference behaviour that copies and amplifies these patterns (Balayn & Gürses, 2021). The result of this process is the ML model, which is a set of

mathematical equations with parameters learned from available data, using the algorithm. This ML model can now be used to make inferences on new data, using the patterns learned from the training data (Balayn & Gürses, 2021). Data refers to the set of numerical information used for executing the algorithm (Balayn & Gürses, 2021).

Further, a distinction is made between an ML model and ML system. Where an ML model is the main artefact explained above that can be used to make inferences on new data, additional components are required to ultimately make such inferences (Balayn & Gürses, 2021). For example, these new data points need to be pre-processed before used as input to the ML model, and need to be post-processed before presented to the user (Balayn & Gürses, 2021). Hence, the ML system refers to the larger technical system consisting of the ML model and additional components needed to make inferences.

1.2.2 ML lifecycle and MLOps

The creation of ML systems comprehends several activities that can roughly be categorized into three stages: the experimental stage, the deployment stage and the operations stage. The experimental stage contains all the steps from identification of the problem till the final ML model. When the ML model is deployed into the organisation and put into service, this results in the ML system. Once deployed, the performance of the ML system should constantly be monitored (Alla & Adari, 2021). If the performance is lacking, the whole process may need to be repeated to update the model, which is very work-intensive (Alla & Adari, 2021). To make deploying and operating ML systems significantly easier, a new approach called MLOps has recently emerged (Alla & Adari, 2021). MLOps is a practice to automate, manage and speed up the ML system's lengthy operationalization (build, test, and release), by integrating DevOps practices into ML (Ruf, Madan, Reich, & Ould-Abdeslam, 2021). DevOps is the go-to methodology for continuous software engineering. In the MLOps field, several technologies and tools are developed for every part of the ML lifecycle (Ruf et al., 2021).

1.2.3 Data driven decision-making

ML systems are used for data driven decision-making (DDDM), In order to understand the different types of DDDM processes that ML systems can impose, a distinction is made between automated decision-making (ADM) and decision support systems (DSS). ADM is defined as "decisions made by technological means without human involvement" (European Commission, 2018, p.8). in ADM systems, the ML system makes the final decision. DSSs on the other hand, are ML systems that inform human decision-makers (Araujo, Helberger, Kruikemeier, & de Vreese, 2020). Within DSSs there is a wide range of possible human-ML system interactions. A human-in-the-loop configuration for example, provides a feedback loop where a human can assess the ML output which can be used to improve the model (Grønsund & Aanestad, 2020). In this master thesis, both types of DDDM processes will be studied.

1.2.4 Sociotechnical complexity of ML systems

Integrating ML into existing social contexts is very complex, and how to effectively and safely do so remains to be contested (Makarius, Mukherjee, Fox, & Fox, 2020). ML systems can impose severe harms, including violations of fundamental rights and economic losses (Balayn & Gürses, 2021). Especially in high stakes domains, such as healthcare and finance, the complexity around potential ML solutions is high. Those domains know many stakeholders, regulations, and decisions made by ML models can have large impact on civilians and society. Introducing ML in a domain makes the ML system part of a larger sociotechnical system. A sociotechnical system contains physical-technical elements and networks of independent actors (de Bruijn & Herder, 2009).

At the same time, harms imposed by ML systems are primarily characterised as 'bias' or technical flaws in the design of the technical system, which leads to a focus on technical solutions (Balayn & Gürses, 2021). This way, a complex sociotechnical problem is narrowed down to a problem in the technical design of AI systems, and thus in the hands of technology companies and technical stakeholders (Balayn & Gürses, 2021). Technology companies then can freely decide on sociotechnical considerations. However, problems such as discrimination cannot be tackled only by technology specialists, but require a more holistic specification and evaluation of ML systems (Balayn & Gürses, 2021).

1.2.5 Sociotechnical ML systems perspective on ML systems

As ML systems are deployed in social contexts, it is vital to consider how the ML system is interlaced with that context (Selbst et al., 2019). To illustrate this, the differences of the concepts "output" and "outcome" are important to understand. The output refers to the inferences ML systems make on new data (Balayn & Gürses, 2021). An ML system is in turn used in a social context, where these outputs impact stakeholders. This impact is defined as the outcome. The outcome can differ from the output, for example when human decision-makers make the final decision, using the ML output. This makes clear that an ML system does not operate in isolation, but is part of a sociotechnical system, consisting of physical-technical elements and networks of independent actors (de Bruijn & Herder, 2009). These components need to be part of analysis, in order to design and operate an ethical and responsible decision-making process (Araujo et al., 2020; Selbst et al., 2019; Green & Chen, 2019b).

The sociotechnical system in this master thesis is defined as a sociotechnical ML system: "a system consisting of technical artefacts, human agents and institutions, in which a machine-based subsystem influences its real or virtual environment by automating, supporting or augmenting decision-making" (Dobbe, 2022, p.3). In this thesis, the scope is limited to ML use cases, while being aware that the machine-based subsystem may vary in terms of its internal model, from rule-based to expert systems to statistical learning models (Dobbe, 2022). It is kept in mind that it should not be predefined that ML is the solution to every problem.

This sociotechnical ML system's view is inspired by the field of "AI safety", which is an "interdisciplinary study of how to build systems that are aligned with the structure of human values, in particular those of stakeholders whom the system is meant to serve" (Dobbe et al., 2021, p.3). Values such as fairness are emergent properties, that arise from the interactions between the system components mentioned before (Dobbe et al., 2021). Therefore, it is important to be aware that a sociotechnical ML system and its components' interactions will change over time, for example through changing behaviours, or a changing environment (Dobbe, 2022).

In Figure 4.1, a visualisation of the scope for this research is presented. The sociotechnical ML system is the centre of analysis for the research. This contains the ML model, ML system, and the IT infrastructure the ML system is part of. Further, institutions, stakeholders and their behaviour, the decision-making process an ML system becomes part of and the outcome of the sociotechnical ML system are part of analysis. To keep this thesis manageable within the required timeframe, ML used for other automation purposes than decision-making, broader organisational changes, transformations and impact on other decision-making processes inferred by the introduction of the sociotechnical ML system are considered outside the scope.

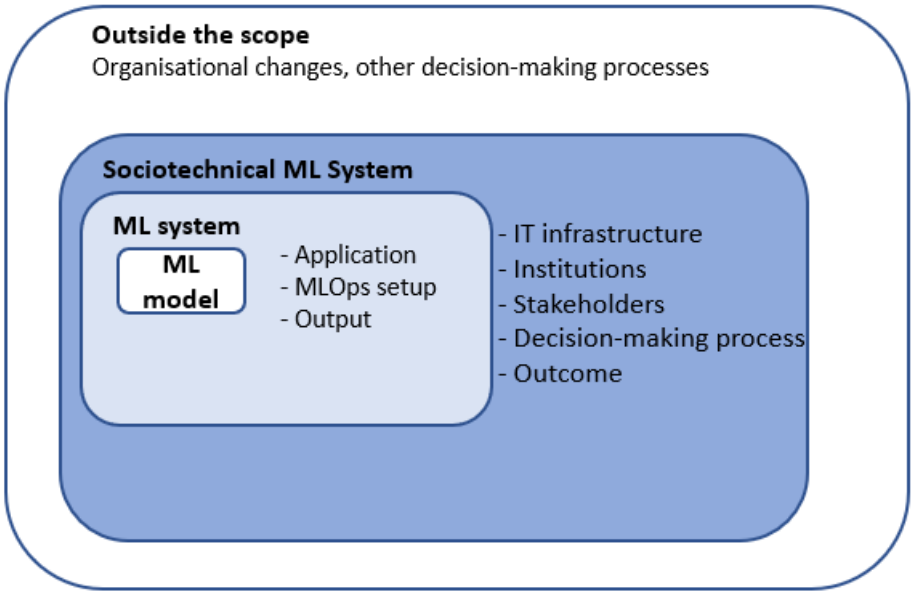

Figure 1.1: Scope of the research

1.2.6 Introduction to Machine Learning in the financial sector

To gain an in-depth understanding of the specification, development and operations of ML in high stakes domains, the high stakes financial domain is selected for the empirical part of this research. Within the financial domain, two ML use cases within two different banks are subject to get an understanding of the development and operations of ML use cases in practice. The empirical insights from these use cases are used to complement academic literature on sociotechnical systems and ML. Those are complementary because there are some key differences when dealing with ML in academics and production (Zhao, 2021). ML in research is mainly focused on trying to improve the ML models, or solving scientific problems, while ML projects in practice are more about engineering (Zhao, 2021). In practice, it is not only about implementation of ML models, but also about integrating the ML system into an existing sociotechnical context.

The financial domain is selected because it is a domain that is held to a higher societal standard than many other domains, because trust in financial institutions is essential for an effective financial system (van der Burgt, 2019). When ML is used in managing financial transactions, relatively small incidents could have major consequences for citizens and society (van der Burgt, 2019). This high impact that ML could have makes the financial domain relevant to study in this master thesis.

The choice to include insights from use cases in two banks is based on the different characteristics, which leads to a more comprehensive understanding of ML within the financial sector than if only one bank would be subject to this research. Both banks are corporate financial organisations, but they have a different level of maturity regarding ML. Bank A has multiple data science teams, developing their own models and having experienced the complexity of integrating them into the organisation. Bank B has a 5-10 FTE marketing data team, which does not develop ML models in-house. The marketing data team has outsourced the development of ML to an external party, of which one first model is currently used in production. Interviews with stakeholders within and associated with the banks provide insights in the domain and organisation specific current ML practices and sociotechnical challenges.

1.3 Knowledge gaps

ML is rapidly adopted in high stake domains as health care and finance, and valuable in those sectors to improve predictions, optimising operations and resource allocation, and personalising service delivery (Dobbe et al., 2021; European Commission, 2021). While the potential of ML is clear, many organisations fail to get the anticipated benefits (Makarius et al., 2020). Moreover, ML systems may cause harms, such as the violation of fundamental rights, economic losses and physical dangers (Balayn & Gürses, 2021; Dobbe et al., 2021).

In response, many reports and statements about how ML should be governed to respect fundamental rights have been published lately (Dobbe et al., 2021). An example is the Assessment List for Trustworthy Artificial Intelligence of the EU, which prescribes seven requirements for AI systems (High Level Expert Group on Artificial Intelligence EU, 2020). Requirements are for example Human Agency and Oversight, Technical Robustness and Safety, and Transparency. Furthermore, Floridi and Cowls (2019) propose a framework consisting of five principles for AI, for example Non-Maleficence, Autonomy, and Justice. Lastly, the DNB initiated six principles for the use of AI in the financial sector, which are soundness, accountability, ethics, skills, and transparency (van der Burgt, 2019). As summarized by Mittelstadt (2019), AI Ethics initiatives have mainly produced high-level principles and vague statements, but provide few specific recommendations that can be used in the practice of developing ML systems. Although these proposals illustrate what it is organisations need to strive for when dealing with ethical, legal and societal implications of AI systems, there is little research on how to consolidate these high-level principles and requirements in the specification of ML systems in practice (Dobbe, 2022).

At the same time, engineering and computer science fields tend to formulate problems and their solutions in technical terms (Dobbe et al., 2021). This is also seen in ML, where MLOps is seen as a potential solution to get ML models actually deployed into organisations, and debiasing as a technical solution to prevent bias in ML systems (Makarius et al., 2020; Alla & Adari, 2021; Balayn & Gürses, 2021). Much of current work in practice pursues technological solutions and metrics to quantify ethical concepts such as fairness (Mittelstadt, 2019).

However, for ML to be adopted in high stakes domains in a safe and effective manner, a more comprehensive lens is required, as sociotechnical complexity and normative stakes are engaged (Dobbe et al., 2021). The need for a sociotechnical systems view on ML is recognised lately by several scholars (Sendak et al., 2020; Makarius et al., 2020; Mateescu & Elish, 2019; Green & Chen, 2019b; Dobbe et al., 2021; Balayn & Gürses, 2021). Dobbe et al. (2021) argue that a sociotechnical ML systems lens can explain how vulnerabilities in sociotechnical ML systems originate from different components and interactions (Dobbe et al., 2021). Systems can not be safeguarded by addressing technical vulnerabilities alone, but need a sociotechnical ML systems view to identify a broader set of vulnerabilities that can lead to harm, if left unaddressed (Dobbe et al., 2021). This leads to the first knowledge gap: A comprehensive overview of what vulnerabilities can emerge in sociotechnical ML systems is lacking, as well as theory that synthesises in which sociotechnical ML system's dimensions they emerge.

This thesis aims at developing a theoretical framework that gives insights in the dimensions in which vulnerabilities emerge, which can form the scientific foundation for sociotechnical ML system specification.

Dobbe et al. (2021) applied a sociotechnical system view in their preliminary work, which proposes AI development as the following cybernetic practices: Sociotechnical Specification, Featurization, Optimization, and Integration. They centre the need for Sociotechnical Specification, which aims to facilitate the different interests in understanding the situation that might benefit from an ML solution (Dobbe et al., 2021). They call for research into particular development domains, to be able to understand how the cybernetic practices should be operationalized in practice (Dobbe et al., 2021). Keeping up with practical development of ML systems is necessary for sociotechnical AI scholarship, as research should be aligned with the rapidly evolving domain, to address the right questions. Mittelstadt (2019) call for the development of guidelines and an empirical knowledge base to understand the impact of the development of ML solutions. Multidisciplinary bottom-up research should be supported by access to practical ML development settings (Mittelstadt, 2019). This leads to the second knowledge gap: There is little insight in how is dealt with sociotechnical ML systems' dimensions of vulnerabilities in the development and use of ML systems in practical settings.

This thesis aims to fill this gap by gathering empirical data from ML development and use in the high-stakes financial domain. These insights can identify best practices on dealing with sociotechnical systems' dimensions, and provide insights in the main challenges in development and use of ML systems, seen from a sociotechnical ML system's perspective. To gather those insights, semi-structured interviews with a variety of stakeholders with different roles involved in two different ML use cases within two different banks are interviewed. To enrich the insights, the perspectives of civil society organisations and regulatory bodies are included by means of semi-structured interviews too.

Ultimately, filling these knowledge gaps provides the insights needed for the design cycle of this research, which aims to develop a sociotechnical guide that can guide the specification of ML systems in practice.

1.4 Research objective and main research question

The research objective is to design pragmatic sociotechnical guidelines that can be used as a starting point in the development of ML use cases. The guidelines will systematically include a sociotechnical ML systems perspective to future use cases, and can be used to re-evaluate existing use cases. The guidelines widen the often technocentric approach of ML practitioners in ML development projects and contribute to an effective and safe development and use of sociotechnical ML systems. The guidelines centre the sociotechnical specification of sociotechnical ML systems. This leads to the following overall research question of this master thesis:

What guidelines should be followed in ML practice to establish a sociotechnical ML systems view in the specification of Machine Learning systems?

The contributions of the research are as follows: First, the research tries to expand the literature on sociotechnical ML development and operations, providing insights in vulnerabilities that emerge in sociotechnical ML systems and how it can synthesised in sociotechnical ML system's

dimensions, augmented with insights from empirical data. Second, it provides insights to the company Deeploy, where the graduation internship is pursued. Currently, Deeploy is developing a platform in which ML models can be deployed in a manageable, accountable and explainable way (Deeploy, n.d.). As Deeploy is a technical tool that supports organisations to effectively deploy ML models, it does not fully grasp the larger sociotechnical ML system the ML models become part of when they are deployed. To Deeploy, the insights of the research provide a comprehensive perspective on the sociotechnical ML system its platform can play a role in. The research provides a starting point for Deeploy for the total design of ML operations, as well as implications for the technical platform itself. Lastly, the research provides a practical starting point for organisations that want to develop and/or use ML systems for their decision-making processes, or a way to re-evaluate sociotechnical ML systems that are already used in practice.

1.5 Link with the CoSEM program

The proposed research is part of obtaining a master degree in Complex Systems Engineering and Management (CoSEM). A CoSEM master thesis is focused on designing in sociotechnical systems. Research in the ML field nowadays often focuses on its technical functioning and technological challenges, but recent publications point out that there is a need for a more comprehensive sociotechnical systems view to realise effective and safe integration of ML into organisations (Makarius et al., 2020; Green & Chen, 2019b; Sendak et al., 2020; Dobbe et al., 2021). This research responds to this need by taking on a sociotechnical ML systems perspective to design guidelines for ML practice, which seamlessly fits the CoSEM program.

1.6 Thesis outline

This report starts with the research approach and methods, outlined in Chapter 2. Then, the knowledge base is laid out in Chapter 3 and Chapter 4, where the first creates an understanding of the ML lifecycle and the latter constructs the theoretical framework for this research. The empirical part of this research is presented in Chapter 5, Chapter 6, and Chapter 7. Finally, all gathered insights lead to the design of a sociotechnical guide in Chapter 8. The research is concluded with the conclusion & discussion in Chapter 9.

Chapter 2

Research approach and methods

This chapter presents the research approach used for this master thesis. Further, the research phases and associated research questions are introduced. Finally, the research methods are presented.

2.1 Selection of the research strategy

To answer the main research question, a design science research (DSR) approach is used. In DSR, the aim is to create novel artefacts to be used by people in order to solve a practical problem (Johannesson & Perjons, 2014). DSR is aimed to produce and communicate knowledge that is of general interest; it should contribute to a global practice as well as to a scientific body of knowledge. To arrive at the design of an artefact for the global interest, a researcher can first design an artefact for a specific problem in a local practice, after which the experience and knowledge can be distilled to inform a general solution (Johannesson & Perjons, 2014). This research strategy is followed in this master thesis project. State-of-the-art literature on specification within the ML development, deployment and operations as well as vulnerabilities that can be present will provide the scientific basis to the design science research project. The local practice of two ML use cases within different financial banks is used to distil empirical data for the design science project. These data are combined to design the artefact. The artefact in turn contributes to improving the local practice, is generalisable to a wider spectrum of practices, so that the artefact contributes to the scientific body of knowledge and to a global practice of ML development and operations. This process is shown in Figure 2.1 below.

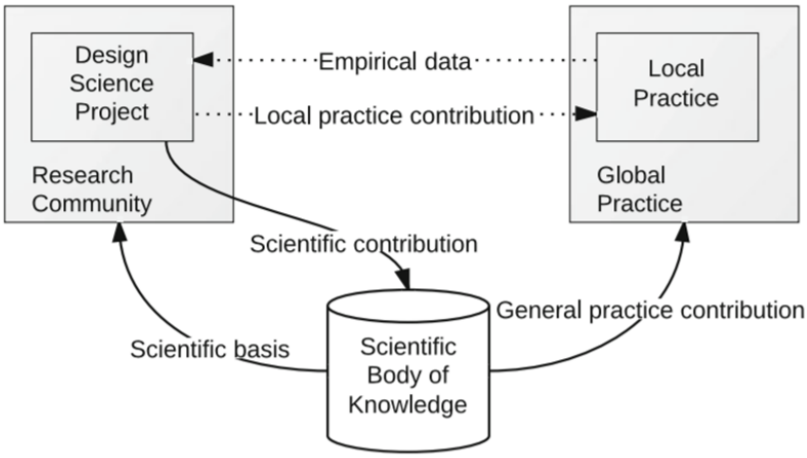

Figure 2.1: DSR of
Johannesson and Perjons (2014)

2.2 The artefact to be designed

There are four types of artefacts constituting the possible outputs of DSR: constructs, models, methods, and instantiations (March & Smith, 1995). The output for this research project is a method. According to Johannesson and Perjons (2014, p. 29), "a method expresses prescriptive knowledge by defining guidelines and processes for how to solve problems and achieve goals. In particular, a method can prescribe how to create artefacts". Descriptive knowledge derived from the scientific knowledge base and the use cases within the local practice will provide directions to the design of the method. The method will prescribe how to specify ML use cases throughout the ML lifecycle. The method will consist of guidelines and a high-level sociotechnical ML lifecycle process to be used as a starting point for ML use cases.

2.3 Design Science Cycles

DSR literature provides several frameworks to guide design researchers in the DSR process, e.g. (Hevner, 2007; Johannesson & Perjons, 2014; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). In these papers, several aspects of DSR and the activities that should be performed to arrive at the design of an innovative and original design are introduced. In this research, the three-cycle view developed by Hevner (2007) is being used. This three-cycle view distinguishes three research cycles that should be present in a DSR project. The Relevance Cycle bridges the contextual environment of the research project with the design science activities. The Rigor Cycle connects the design science activities with the existing knowledge base of scientific literature, experience and expertise. The Design Cycle iterates between building and evaluating the to be designed artefact (Hevner, 2007). This view is aligned with the research strategy, as the existing knowledge base as well as the local practice environment provide input to the design cycle. The focus in this research lies on creating an understanding of the knowledge base and the environment, which leads to a first design. According to Hevner (2007), the design should be field-tested within the local practice environment to iteratively improve the design, after which the design can be implemented in the local practice environment. However, due to time limitations and the importance to first create a thorough understanding of the knowledge base and the local practice environment, the research concludes with a first design, and leaves field-testing and iterating up for future research. In Figure 2.2, the three-cycle view that is followed in this research is presented, including the different sub-questions aligned with the cycles.

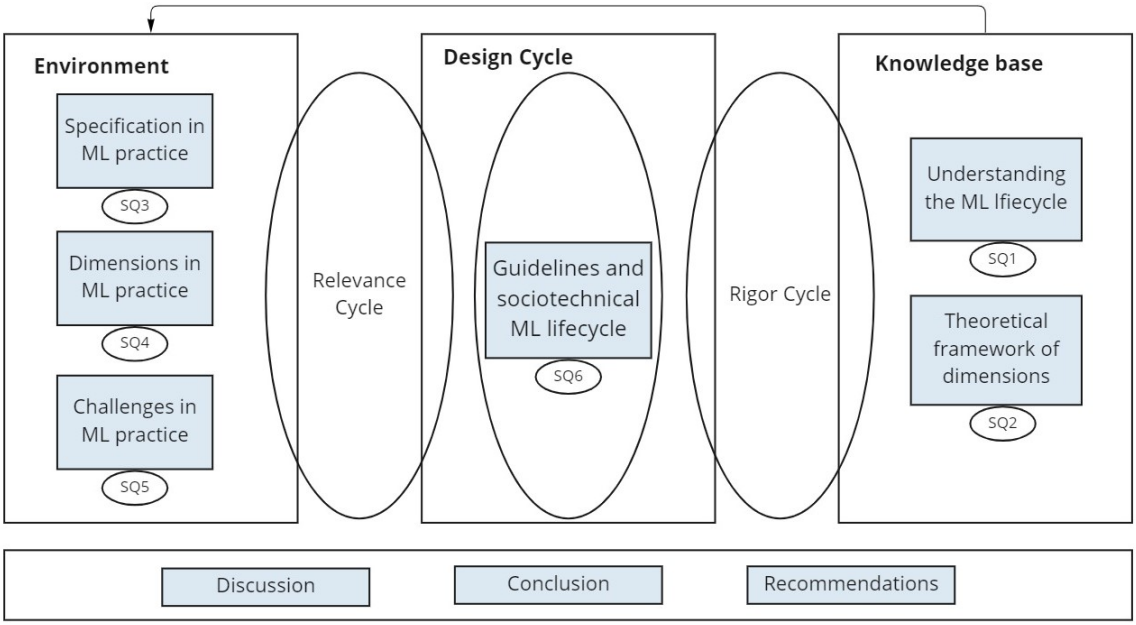

Figure 2.2: Three-cycle design approach for this research adapted from Hevner (2007)

2.4 Research phases and sub-questions

The research consists of three subsequent phases, aligned with the Design Science Cycles. In each phase, one or more sub-questions are answered. These together answer the main research question.

2.4.1 Research phase 1: Understanding the ML lifecycle and dimensions in which vulnerabilities arise

Research phase 1 focuses on getting a thorough understanding of the knowledge base. First, a thorough understanding of what it entails to develop, deploy and operate ML use cases is reached, by means of desk research. This results in a general overview of what the ML lifecycle looks like. Further, the role of the recent emergence of the MLOps practice and the Deeploy platform within the ML lifecycle is presented. Lastly, the limitations of the general ML lifecycle and MLOps based on literature on sociotechnical ML systems are presented. This results in answering the first sub-question:

1. What does the Machine Learning lifecycle look like in general, and what limitations can be identified from a sociotechnical ML systems perspective?

Knowing the limitations of the current ML lifecycle and MLOps, provides a basis for directing the research towards a sociotechnical approach to apply in the ML lifecycle. This starts with the introduction of sociotechnical specification based on existing literature, after which a literature review on vulnerabilities is performed. Based on the vulnerabilities resulting from the literature review, a theoretical framework is constructed to synthesize the vulnerabilities in sociotechnical ML system's dimensions in which vulnerabilities arise. This results in answering the second sub-question:

2. In which dimensions do vulnerabilities arise, seen from a sociotechnical Machine Learning systems perspective?

2.4.2 Research phase 2: Specification, addressing of dimensions and identify main sociotechnical challenges in practice

The second research phase builds upon the knowledge base laid out in the first research phase. The gathered insights are used to enter the field and gather data from the local practice environment. The local practice subject for this research is mainly centred in the financial sector, diving into two ML use cases that have been developed in two different banks. Interviews with different stakeholders with different roles are conducted to gather insight from the local practice environment. First, the gathered insights in the knowledge base on sociotechnical specification are projected onto the two use cases, to answer the third sub-question:

3. What does specification in the Machine Learning lifecycle look like in practice, based on two use cases in the financial sector?

Subsequently, an analysis is performed to gather insight on how the dimensions formulated in sub-question 2 play a role in the ML lifecycle in practice. To this extent, data is gathered conducting the interviews with stakeholders involved in the use cases. Additionally, more general empirical insights on the knowledge base from the perspectives of civil society organisations and regulatory bodies are gathered by conducting interviews with their representatives. This leads to answering the fourth sub-question:

4. To what extent are sociotechnical dimensions addressed in practice, based on use case specific and general insights?

The next step is to perform an inductive analysis to formulate the main challenges identified along the dimensions and informed by additional interview data. This will result in answering the fifth sub-question:

5. What are the main challenges identified in ML practice, seen from a sociotechnical ML system perspective?

2.4.3 Research phase 3: Design a sociotechnical guide

The last step in the research reflects the design cycle of the DSR. In this step, all gather insights inform the creative process of the researcher to design a sociotechnical guide for ML practice. This results in answering the last sub-question:

6. What guidelines can guide ML practice in the sociotechnical specification of sociotechnical ML systems?

2.5 Research method per sub-question

Sub-question 1: (Grey) literature review and document analysis
To create an understanding of the general ML lifecycle and the emergence of MLOps, reviewing (grey) literature, document analysis are used as methods to answer this sub-question. As MLOps is a quite new topic, there is little academic literature available on MLOps (Zhao, 2021). Therefore, the academic literature used for this sub-question is complemented with grey literature, such as commercial books. To identify the stakeholders and their roles, document analysis has been performed. Document analysis has also been performed to clarify different relevant concepts. Last, to validate the general ML lifecycle, discussions with the external supervisor for this research are held.

Sub-question 2: Integrative literature review and Grounded theory building
To answer the second sub-question, an integrative literature review is used as a method. An integrative literature reviews, critiques, and synthesizes representative literature on a topic in a way that new perspectives and frameworks on the topic are generated (Torraco, 2005). In this research, the integrative literature review focuses on synthesizing representative literature on vulnerabilities that emerge in sociotechnical ML systems, to be able to construct a theoretical framework. Since researching vulnerabilities in sociotechnical ML systems is a new topic, the purpose of the literature review is rather to create an initial theoretical framework, rather than review old frameworks (Snyder, 2019). The integrative literature requires a more creative collection of data, as the purpose is to combine perspectives and insights from different fields and research traditions, instead of trying to cover all articles ever published on a topic (Snyder, 2019). A limitation of the integrative literature review method is that it can be difficult to provide transparency on the review process, as no pre-specified inclusion criteria are used, as is the case in a systematic literature review (Snyder, 2019). As it is important is to provide transparency, an overview into the literature selected, the main themes, and databases used are beneficial (Torraco, 2005). The database used is Scopus. The other information is provided in Appendix A.

To arrive at the theoretical framework, the Grounded Theory building method has been used. This method follows an inductive approach in order to generate or discover theory (Torraco, 2002). The theory evolves through continuous interplay between analysis and data collection (Torraco, 2002). As a result, Grounded Theory building allows new theoretical understandings to emerge from the data (Torraco, 2002). The data used to construct the theoretical framework is collected by the integrative literature review on vulnerabilities. The researcher applies conceptual thinking to synthesize the data to arrive at the theoretical framework.

Sub-question 3, 4, and 5: Semi-structured interviews
To answer the third, fourth, and fifth sub-question, semi-structured interviews are used as a research method. Semi-structured interviews are the most common qualitative research method (Qu & Dumay, 2011). Questioning must be prepared and guided by identified themes in a consistent and systematic manner (Qu & Dumay, 2011). Because the research involves semi-structured interviews in two rounds, two separate interview protocols are created, one for the interviews with stakeholders involved in the ML use cases and one for the interviews conducted with representatives of civil society organisations and regulatory bodies. Important is that questions asked must be comprehensible to the interviewee (Qu & Dumay, 2011). This is taken into account in the preparation of every single interview, as a variety of stakeholders with different roles are interviews. Therefore, in each interview, appropriate questions to the specific interviewee will asked, while adhering to the identified themes upfront to ensure the same thematic approach is applied during each interview (Qu & Dumay, 2011). The themes are mainly specified based on the knowledge base, to gather empirical insights on the theoretical framework and specification process. A disadvantage of semi-structured interviews is that planning, preparing, conducting, and analysing

the interviews are very time-consuming (Hove & Anda, 2005). Furthermore, the way in which the interview is conducted determines the quality of the collected data, which makes it important that the interviews are carried out carefully (Hove & Anda, 2005). In the research, both deductive and inductive analyses of the interview data are performed to answer the sub-questions. First, a deductive analysis is performed, in which the coding frame has been developed at the beginning of the analysis process (Friese, Soratto, & Pires, 2018). Performing deductive analyses give insight in the specification process in the ML use cases (sub-question 3) and in to what extent the dimensions are addressed in practice (sub-question 4). For the fifth sub-question, an inductive analysis is performed, in which the researcher is not trying to fit data into a pre-existing coding frame (Friese et al., 2018). Rather, the insights along the dimensions and additional interview data are used to identify the main challenges in ML practice.

Sub-question 6: Building upon previous research methods
To answer the sixth sub-question, all insights gathered by the research methods in previous sections are used in a creative design process to develop the sociotechnical guide. As such, answering this sub-question does not need a distinct research method, but is a design process.

2.6 Data gathering and processing

To gather data for the first sub-question, Scopus is used to find literature and Google Scholar is used to find grey literature on the ML Lifecycle and MLOps. Google is used for the document analysis. To gather data for the second sub-question, Scopus is used as a tool to find scientific literature.

For the second, third, and fourth questions, the answers to the interview questions form the data, so the answers have to be recorded and processed systematically. Eighteen interviews have been conducted, using Microsoft Teams as a tool to conduct, record and transcribe the interviews. To systematically analyse the interview transcripts, coding is performed. Coding helps in organising, structuring and retrieving data (Friese et al., 2018). The software ATLAS.ti is used for coding, which is a tool that has been used by professionals and researchers across different fields of knowledge (Friese et al., 2018).

2.7 Research Flow diagram

The flow of conducting the research is presented in the Research Flow diagram in Figure 2.3. The Research Flow diagram presents the three research phases, consisting of research activities leading to answering the sub-questions in each chapter. The research methods are presented in the rights bottom of every chapter. The arrows present the research outputs of research phases that serve as input to subsequent phases. Combining insights from all phases leads to answering the main research question.

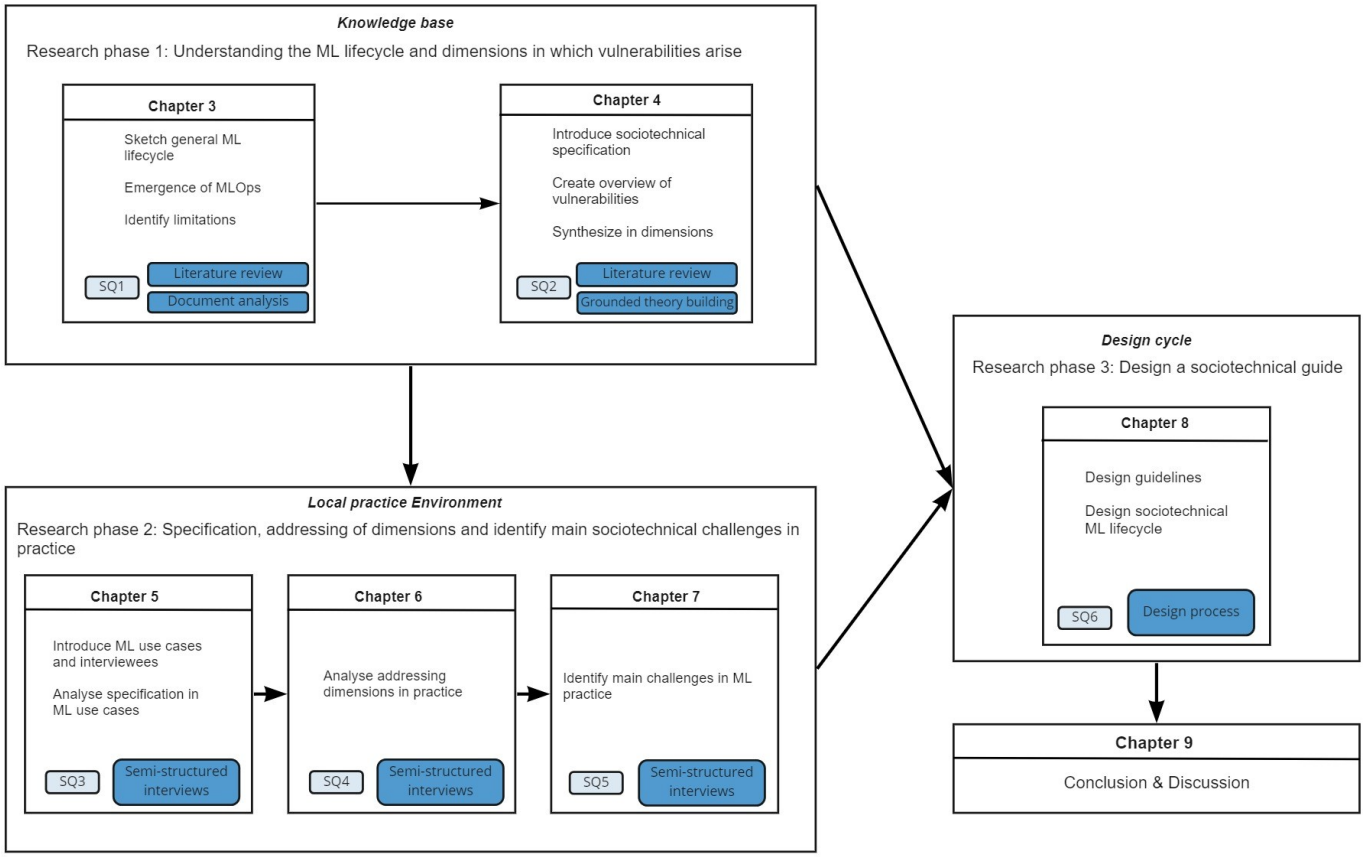

Figure 2.3: Research flow diagram

Chapter 3

The ML lifecycle and its limitations

This chapter contains the first results from the first phase of the research, which entails defining the knowledge base for this research. First, a general understanding of the ML lifecycle and the stakeholders involved is reached by means of reviewing (grey) literature and described in Section 3.1. In Section 3.2, limitations of the ML lifecycle are identified. Further, the recent emergence of the MLOps practice for the ML lifecycle is described in Section 3.3. Lastly, a reflection from a sociotechnical ML systems perspective of MLOps is presented in Section 3.4. This chapter aims to answer the first sub-question:

1. What does the Machine Learning lifecycle look like in general, and what limitations can be identified from a sociotechnical ML systems perspective?

3.1 General design of the ML lifecycle

Combining insights from different literature and practice (N. Jetten, Personal Communication, October 22, 2021), Figure 3.1 presents all the general steps of the ML lifecycle. The ML Lifecycle can roughly be divided into three stages: experimental stage, deployment stage and operations stage. The experimental stage involves all steps that lead to an ML model. The deployment stage includes the steps to finally integrate the model in an organisations' infrastructure, so that it can be used to make predictions and use them in business. The operations stage comprises the monitoring of the model and application. The subsequent paragraphs explain the activities in each stage.

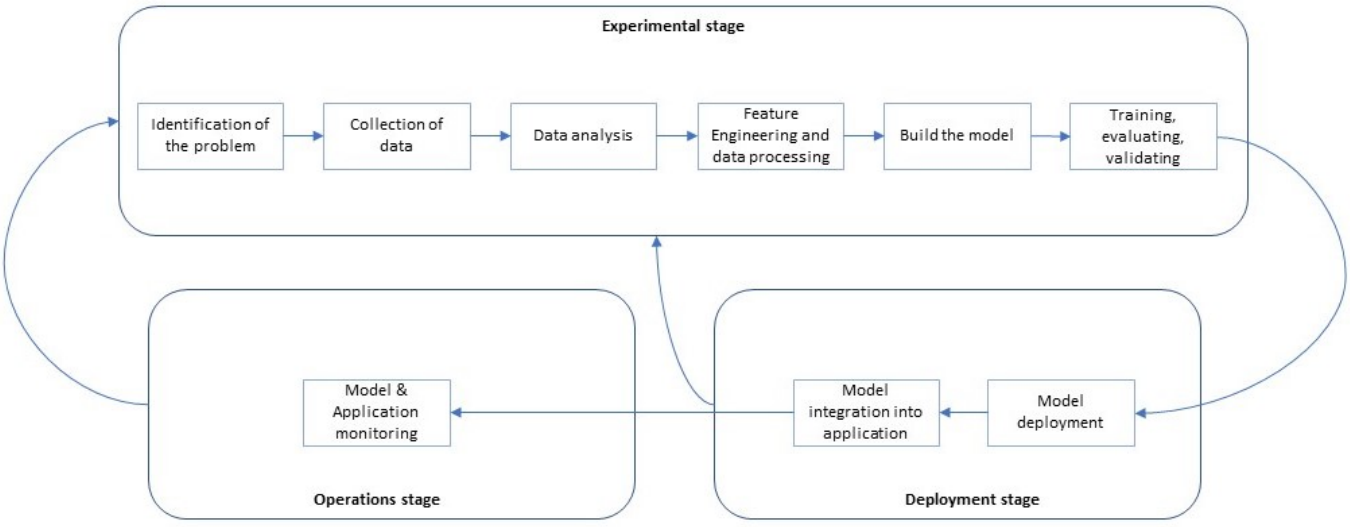

Figure 3.1: ML lifecycle

3.1.1 Experimental stage

The experimental stage begins with the identification of the problem, defining the problem and why it is worth solving (Alla & Adari, 2021). Once the problem is identified, one should collect data, which could be structured data, credit card data for example, or unstructured data, pictures for example. Once the raw data are collected, one should clean, process and format the data, afterwards a data analysis can be performed. Data analysis could be constructing graphs to get an idea of the overall distribution and significant relationships between features (Alla & Adari, 2021). The next step is feature engineering and data processing. Feature engineering is the creation of new features from the combination of several existing ones, which is done to give the model a deeper context, so it can learn the task better. Data processing entails preparing the data so that they can be passed into a model. This step also involves creating subsets of the initial data into a training data set, a testing data set, and a validation data set. The training data set is the data the model learns from, the testing data set contains data to evaluate the model's performance on, and the validation data set is used to select a model or tune the model's hyperparameters to accomplish a better performance of the model (Alla & Adari, 2021). Afterwards, a model is built. Lastly, the training, evaluating and validating step is performed (Alla & Adari, 2021). First, the model is trained, it learns how to perform the task for which the model is developed. After training, either the evaluation step or the validation step follows. In the evaluation step, the model's performance is measured by metrics such as accuracy, precision, recall, by inputting data that the model has never seen before. What the validation step includes depends on the context. Examples for which validation is used are selection of the model architecture, selecting the best model or tuning hyperparameters. The experimental stage is an iterative process, as several models using different algorithms might be trained and tested, feature selections might be adapted and the final model's hyperparameters may be tuned (Treveil, 2020). Once the experimental stage is completed, the model is ready to make predictions.

3.1.2 Deployment stage

The deployment stage includes the steps that need to be taken in order to integrate the trained model into the software infrastructure, so that its predictions become available to other users or other systems (Paleyes, Urma, & Lawrence, 2020). Within the deployment stage, there are two sub activities: model deployment and integration into organisation. There are two common types of model deployment: model-as-a-service and embedded model (Treveil, 2020). In model-as-a-service, the model is deployed as an Application Programmable Interface (API) on a web server. An API is "a set of programming code that enables data transmission between one software product and another, also containing the terms of this data exchange" (Altexsoft, 2021). In other words, an API allows two software applications to communicate. An ML model deployed as an API makes it possible to use the ML model's predictions, accessed by a web application. An API contains one or more API endpoints, which are the URLs that can be accessed by an application to make a request and receive a response (Curtis, 2020). This API endpoint can respond to requests in real time, meaning input data for the ML model to make a prediction on can be sent to the API endpoint, The API sends the input data to the ML model, the model uses the input data to make a prediction, and the resulting prediction is sent back as a response. An embedded model means that a model is packaged into an application that is published (Treveil, 2020). An embedded model does not provide predictions real-time, but in batches, typically on a recurring schedule on data that became available in that batch (e.g. hourly, daily).

The second step is to integrate the deployed model into the organisation. Now that it is possible to use the deployed model to make predictions on input data, business experts or customers should be able to provide input data. This requires that the deployed model is integrated in an external application, for example a website. An example of the end result in the case of model-as-a-service would be a user that fills in data on a website, which is being sent to the endpoint of the deployed model to make a prediction, and the prediction as output presented on the website.

3.1.3 Operations stage

Once the model is actually used in the organisation, the whole ML system should be monitored (Ruf et al., 2021). This means model and application performance, monitoring infrastructural circumstances and the data utilized by the model. Monitoring is essential to ensure the model keeps performing well and behaves as expected. Also, monitoring is required so that the organisation

has a precise knowledge of how broadly each model is used (Treveil, 2020). However, the research community is in the early stages of understanding what are the key metrics of data and models to monitor and how to alert on them (Paleyes et al., 2020).

3.1.4 Iterative nature of the ML lifecycle

Important to understand is that the ML lifecycle has an iterative nature, and is not a static end-to-end workflow (Zhao, 2021). In each of the steps, issues might be found that lead to the necessity to modify other stages in the ML lifecycle. This is presented by the arrows flowing from the Deployment Stage to the Experimental stage and from the Operations stage to the Experimental stage in Figure 3.2. Therefore, a process is needed to track and version the data, code, model, results and each experiment, so that changes can be made easily and reproducibility of all stages is ensured (Zhao, 2021).

3.1.5 Stakeholders in the ML lifecycle

There are a variety of stakeholders that may be involved in the ML lifecycle, both inside an organisation and outside. Depending on how an organisation is organised, some stakeholders take on multiple roles. Table 3.1 present the roles that can be taken in the ML lifecycle. The table presents stakeholders directly involved in the development of an ML system, such as the data engineer and data scientist. The directly involved stakeholders can be employees of the bank, when ML systems are developed in house, as well as external parties, if (part of) ML system development is outsources. Besides, stakeholders that are not directly involved in the development or use of ML systems, but have an interest in ML systems are presented in Table 3.2. Those stakeholders are for example regulators such as DNB and the Autoriteit Persoonsgegevens, or civil society organisations.

Table 3.1: Direct stakeholder roles in the ML lifecycle

Stakeholder role	**Role description**	**ML lifecycle activities involved**
Data provider	Provides the data sets needed for the ML model (Li et al., 2018)	Collection of data
Data engineer	Responsible for transforming unprocessed data sets into interpretive data for the ML model (Ruf et al., 2021)	Collection of data, data analysis
Data scientist	Builds and delivers models that address the requirements provided by the domain expert (Treveil, 2020)	Identification of the problem, Feature engineering and data processing, build the model, training, evaluating, and validating
Software engineer	Integrates ML models in the company's applications and systems (Treveil, 2020)	Model deployment stage
ML engineer/MLOps engineer	Realises model operationalization, monitoring and continuous integration and delivery of the ML lifecycle (Ruf et al., 2021)	Model deployment, Operations stage
ML architect	Ensure a scalable and flexible environment for ML model pipelines and introduce new technologies (Treveil, 2020)	All stages
Data steward	Supervises data quality and data governance (Ruf et al., 2021)	Data collection, data analysis, feature engineering and data processing
Problem owner	Initiator of the ML project, provides requirements for the ML model and validates results (Ruf et al., 2021)	Identification of the problem, model & application monitoring
Subject matter expert	Provides requirements for the ML model and validates results (Ruf et al., 2021)	Identification of the problem, evaluating, validating
End-user	The person consuming the output of an ML model or making a decision based on the model output (Bhatt et al., 2020)	Operations stage
Compliance officer	Ensure compliance with external and internal requirements and regulations (Treveil, 2020)	Deployment stage, Operations stage
Risk manager/auditor	Minimize overall risk to the company caused by ML models in production (Treveil, 2020)	Deployment stage, Operations stage
Board level representative	Carries the ultimate responsibility for activities within the organisation	Deployment stage, Operations stage
Bank customer	Outcome of the ML system impacts the bank's customers	Operations stage

Table 3.2: External stakeholder roles in the ML lifecycle

Stakeholder role	Role description
De Nederlandsche Bank (DNB)	Provides guidelines for the use of AI in the financial sector
The Dutch Authority for Financial Markets (AFM)	Regulatory body on financial markets
Authority for Consumers and Markets (ACM)	Regulatory body that ensures fair competition in financial markets between businesses, and protects consumer interests
Autoriteit Persoonsgegevens (DPA)	Regulatory body on AI and algorithms that use personal data, monitors GDPR compliance
Nederlandse Vereniging van Banken (NVB)	Promotes interests of Dutch banks and foreign banks active in The Netherlands
Civil Society organisations	Non-profit organisations that address the impact of ML on civil rights
European Commission	Formulates EU regulations on AI and ML
Technology tools providers	Provide tools to support (a part of) the ML lifecycle

The following figure provides an overview of the stakeholders involved in which activity of the ML lifecycle.

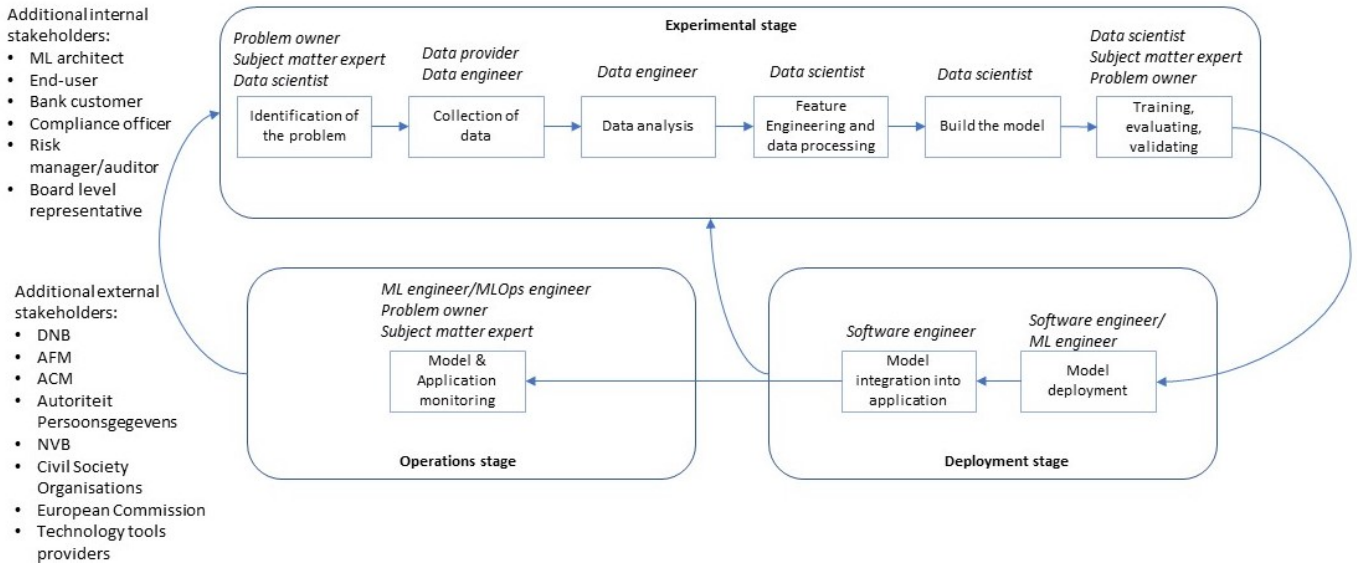

Figure 3.2: ML lifecycle with stakeholders involved

3.2 Limitations of a technical view in the ML lifecycle

The ML lifecycle presents a technical view on the development of an ML system. The technical view focuses on hardware, software, algorithms, mechanical linkages and inputs/outputs (Alter, 2010). This is a narrow view, as the ML system becomes part of a larger sociotechnical ML system, which also include other technical systems, stakeholders, decision-making processes, institutions and the final outcomes, as presented in Figure 4.1. The ML system is a subsystem of the sociotechnical ML system, which consist of technological and social subsystems. This makes clear that there is a gap between the technical conceptual framework, the ML lifecycle, and the needed sociotechnical conceptualisation of the sociotechnical ML system (Alter, 2010).

The following illustrates that taking a technical view is a too narrow view to incorporate social concepts in development of a technical system. ML models are designed and built to achieve specific goals and performance metrics, such as accuracy, precision, and recall. Once an ML system is put to practice, the decision-making process that emerges must be ethical and responsible (Green

& Chen, 2019b). Resulting harms that may be imposed by ML systems are primarily characterised as bias or technical flaws in the design of the system, which leads to a focus on technical solutions (Balayn & Gürses, 2021).

Using this technical view, researchers try to develop ethical and responsible ML models, for example using fairness as a property of the ML model, so that it can be used as a performance metric (Selbst et al., 2019). Those researchers see biases in ML as the problem that lead to unfair systems. Bias in ML is an issue that is either observed in the outputs of the ML models or in the internal representations that the models rely on. If a data set is biased, the outputs of the ML model could be biased too, which might be harmful in discriminating ways when the models are applied in the real world (Balayn & Gürses, 2021). To take away these biases, debiasing is seen by researchers and practitioners as a method to do so. Debiasing is organised as follows; first, a "fairness metric" needs to be selected, which, in theory, reflects the ideal outputs of the ML model when it is "unbiased". Then, according to the fairness metric, a debiasing method is chosen and applied to either the data, the ML algorithm or its outputs. The debiasing method tends to transform one of those, to fulfil the selected metric (Balayn & Gürses, 2021). However, it is technically impossible to optimize for more than one fairness metric at the same time, which makes the potential of debiasing to create fair ML models limited. Moreover, computer science researchers develop debiasing methods centred solely around the inputs and outputs of the ML model. However, the output of the model is only the inference the ML model makes on new data, and does not consider the more relevant outcome of the ML system, which consists of the output and its negative or positive impact on different stakeholders (Balayn & Gürses, 2021). This way, a complex sociotechnical problem is narrowed down to a problem in the technical design of ML systems, and thus in the hands of technology companies and technical stakeholders (Balayn & Gürses, 2021). Technology companies then can freely decide on sociotechnical considerations. However, problems as discrimination cannot be tackled only by technology specialists, but require a more holistic evaluation of ML systems (Balayn & Gürses, 2021). As Selbst et al. (2019) state; fairness and justice are properties of social and legal systems, not of the technical subsystems within.

After the development of the model itself, deploying the model and maintaining it in operation are the subsequent steps in the ML lifecycle. However, deploying models manually is a massive undertaking, and manually maintaining them is work-intensive. Moreover, when a new model version is developed or the model is updated with new training data, reintegration into the application has to be performed manually every time. As a result, most models simply never make it past the model development phase (Alla & Adari, 2021). A new practice called MLOps has recently emerged to address this challenge, which is described in Section 3.3.

3.3 Emergence of MLOps practices in the ML lifecycle

3.3.1 MLOps as a method to address ML lifecycle challenges

MLOps is a relatively new discipline that emerged due to several reasons. Most traditional organisations are relatively new to the development and deployment of ML models. At first, the number of models and the dependency on these models on a company-wide level may be limited (Treveil, 2020). Over time, ML models are developed for more applications, and also increasingly critical decision-making processes, which makes managing their risks more important (Treveil, 2020). Deploying models manually is a massive undertaking, and even if the model is deployed, manually maintaining them is work-intensive, and every new update requires reintegration into the application. Manual ML workflows are a source of high technical debt. Where developing the ML model is relatively fast, maintaining ML systems over time is difficult and expensive (Sculley et al., 2015). This is the case because unlike standard software, ML systems consist of code and data, which make the systems much harder to maintain in the long rung (Ruf et al., 2021). Furthermore, different stakeholders are involved in the ML lifecycle, who all use different tooling and do not share the same fundamental skills for a baseline of communication (Ruf et al., 2021; Treveil, 2020). Moreover, now that organisations develop more and more models, with increasingly critical applications, there is a need to track and version each component, data, code, results and each experiment in the ML lifecycle, to keep control, easily fix changes and to reproduce any stage in the lifecycle (Zhao, 2021).

MLOps aims to accommodate for these challenges by enabling continuous developing, experimenting and improving models, while keeping track of the data sets, models, metrics, etc. (Zhao,

2021). Summarizing, MLOps is a discipline that enables reliable and efficient ML development, deployment and operations (Ruf et al., 2021).

3.3.2 Fundamentals of MLOps

MLOps is a combination of ML, DevOps and Data Engineering. It is a practice to automate, manage and speed up the ML model's lengthy operationalisation (build, test, and release), by integrating DevOps practices into ML (Ruf et al., 2021). At its core, MLOps is the standardisation and streamlining of ML lifecycle management (Treveil, 2020)
.
DevOps has emerged as the new go-to methodology for continuous software engineering. Traditionally, Development (Dev) and Operations (Ops) had hardly any overlap. Dev was responsible for translating the business idea into code, performed by software engineers and developers, and the Ops was responsible to provide a stable, fast and responsive system to the customer, which was performed by the Operations engineers and IT specialists (Ruf et al., 2021). DevOps practice aims at bridging the gap between those two, emphasizing communication, collaboration, and integration between Software Developers and Operations teams. DevOps core practices are Continuous Integration (CI) and Continuous Delivery (CD). CI is a software practice that focuses on automating the process of integrating code from multiple developers (Ruf et al., 2021). CD serves the goal of ensuring that new features developed become available to the end user as soon as possible. CD practices facilitate this, as software is built in a manner that is always in a production-ready state. An optional practice is Continuous Deployment, in which every change is deployed in production automatically (Ruf et al., 2021).

In the MLOps field, several technologies and tools are developed for every part of the ML lifecycle (Ruf et al., 2021). The general desire in MLOps is to automate the ML lifecycle as far as possible to speed up the deployment and operations processes (Treveil, 2020). MLOps applies CI in the ML model developing process, from preparing the data, tuning parameters, running experiments, to building and validating models (Zhao, 2021). CD is used to automatically deliver processed datasets and trained models from the data scientist to the software engineers/MLOps engineers. Also, Continuous Training (CT) is applied, in which the arrival of new data or decreased model performance trigger automatic model retraining (Zhou, Yu, & Ding, 2020).

To arrive at an MLOps setup, one can make use of automated pipelines. The overall ML lifecycle consists of different activities, that can be organised in different pipelines. A pipeline consists of a series of automated processes, that executes sequential parts of the ML lifecycle (Ruf et al., 2021). An example is a pipeline that executes all the steps it takes from data extraction to the resulting ML model automatically (Valohai, n.d.). The desire of MLOps is to arrive at a setup that consists of automated pipelines, without the barriers of manual testing and integration (Alla & Adari, 2021)

While DevOps principles are used in MLOps, there is a significant difference between the two. That is, deploying software code into production is fundamentally different from deploying ML models into production. While software code is relatively static, ML models are constantly learning and adapting, because the world, and thus data, is always changing. This complexity, including the fact that ML models are made up from both code and data, is making MLOps a new and unique discipline (Treveil, 2020).

3.3.3 MLOps platform Deeploy

Deeploy offers a software platform for organisations to enable explainable, accountable, and trustworthy ML deployments by design (Deeploy, n.d.). Deeploy facilitates explainability by offering methods to make models and their consequential decisions explainable to developers, business experts and customers. Accountability refers to making every step of the decision-making traceable and reproducible. Also, decisions made in the past can be reproduced and explained again. Lastly, Deeploy makes ML deployments trustworthy by enabling interaction between models and humans, by providing methods to give feedback and overrule decisions in order to learn and improve models. Deeploy focuses on the Deployment and the Operations stage of the ML lifecycle. Specifically, the Deeploy software is built for the model deployment and the model & application monitoring activities. The functionalities are explained in the following paragraphs, based on personal communication with Deeploy's technical director (T. Kleinloog, personal communication, October 22, 2021).

Deeploy and the deployment stage

Deeploy requires an organisation to use a version control system for the ML model (e.g. GitLab, GitHub, Bitbucket). A version control system is used to store code scripts used for building an ML model and to store the final model as a model object, as well as logging changes in code and model objects. In a version control system, a repository can be created to store all code and ML models for a project. Notable is that a version control system does not accommodate for tracking experiments in the experimental stage, for which other tools can be used, such as MLFlow (MLFlow, n.d.). A standard version control system does not store the datasets used to train the model either. Data versioning is supported by other tools such as Data Version Control (DVC) which can be used in combination with the versioning control system to accomplish versioning of the entire workflow (DVC, n.d.).
In Deeploy, a path to the project repository is created and presented in Deeploy 'repositories'. Next, in Deeploy 'deployments', specific model objects from the repository can be deployed. The resulting deployment contains an API endpoint and additional services that are required to ensure logging and monitoring of predictions. The API endpoint is created to enable communication with the deployed model from external applications. Deeploy provides code snippets around the API endpoint to smoothen the integration process for an organisation's software engineers. These are templates that make it easier to enter repeating code patterns needed to integrate the API endpoint (Visual Studio Code, n.d.). This way, the API endpoint can easily be integrated in an external application, for example in the organisation's internal back-end system or in a customer facing system, so that business experts or the organisation's customers can interact with the model. The integration of the endpoint in the external application is done by software engineers within the client organisations.

Model and application monitoring stage

The models that are deployed and are used in the client's organisation, can be monitored in Deeploy. Deeploy offers four types of monitoring: technical monitoring, concept drift monitoring, feedback monitoring, and event monitoring.

- Technical monitoring ensures that the model deployment functions as expected. Technical monitoring entails checking whether predictions are created, do errors occur, the activity over time is measured and time needed to create a prediction is measured.
- Concept drift monitoring monitors the concept drift of the model's predictions. Concept drift is a phenomenon where the statistical properties of the target variable, which the model is trying to predict, change over time. This might require the data scientist to retrain the model on a new or expanded dataset, or even to improve the model.
- Feedback monitoring relates to expert feedback on a specific prediction, that can be given in Deeploy. Business experts can either validate or overrule a specific prediction made by a model. Explainability of the prediction is essential for the business expert to understand why and how the model predicted a certain outcome. This explanation is therefore displayed by Deeploy.
- Event monitoring entails the logging of changes in the metadata of a deployment. For example, every deployment in Deeploy must have an owner, and a change of the owner is logged as an event.

Deeploy in relation to MLOps

As defined earlier, MLOps is a method to automate, manage and speed up the ML model's lengthy operationalisation (build, test, and release), by integrating DevOps practices into ML (Ruf et al., 2021). Deeploy does this for a part of the ML lifecycle: the deployment and operations stages. Continuous ML (re-)deployment is enabled by creating the link to the version control system repository and deploying a model with a few clicks in Deeploy. This enables data scientists themselves to deploy their models, without the need of software engineers. Additionally, the API endpoint created and integrated in the organisation's external application does not change when changes to the model deployed are made. This makes quick redeployment possible. For the operations stage, Deeploy provides a way in which the data scientists can monitor their own models, instead of the need for an operations team. Also, in Deeploy, a business expert can interact with the predictions made by the models, by looking at the output and explanation provided and providing feedback.

Other than several scholars, Deeploy's vision is not to automate as much of the ML lifecycle as possible. According to Deeploy, the focus must lie on human control of models and its predictions. Where Alla and Adari (2021) suggest automatic redeployment of an altered model, Deeploy requires a manual click to deploy this new version of the model. As ML models' predictions can have a direct impact on people, Deeploy points out that human oversight is essential. Therefore, Deeploy does not promote automatic retraining of models, but supports the data scientist by providing monitoring possibilities, based on which a decision for the data scientist to retrain the model can be made.

3.4 Reflection on MLOps from a sociotechnical ML systems perspective

MLOps introduces ways to keep ML systems in control over time, using versioning and tracking of data sets, models, metrics and experiments. These functions are desired as they improve accountability and the possibility to maintain and improve the ML system over time. At the same time, it introduced CI, CD, and CT to automate the ML lifecycle as far as possible. These automatic pipelines remove the barriers of manual testing and integration. MLOps arose to address technical challenges in the ML lifecycle, offering technical solutions. However, the ML system is required to not only be integrated in technical applications, but becomes part of a sociotechnical ML system, that consists of both technical systems, that contain technology and processes, and social systems, that consist of people and relationships (de Bruijn & Herder, 2009). As such, automating and speeding up the ML lifecycle does bring advantages from a technical efficiency perspective, but insufficiently considers the impact on the larger sociotechnical ML systems. Several limitations have been identified in the MLOps practice from a sociotechnical ML systems perspective, which are described below.

First, Continuous Integration enables automated model validation, after which the Continuous Delivery pipeline automatically delivers the model to be deployed. The limitation identified here is that model validation does not consider the validation of the model's interaction with its sociotechnical context. As such, a validated new model version in the CI pipeline could be valid from a technical perspective, but could be not meeting requirements of stakeholders in the sociotechnical ML system. This way, new model versions are released that could require changes in e.g. the design of decision-making process or interpretation of the model output by end-users, which are not considered here.

Further, Deeploy proposes human control to prevent new model versions to be automatically deployed as described in the previous paragraph. In Deeploy, a human controller, often a data scientist, must manually deploy a new model version. Whereas the intention is to maintain control over which model version is deployed, a simple click on the button is insufficient to address the potential emergent issues that a new model version could cause in the sociotechnical ML system. Even if the human controller has been given instructions on what activities to perform before deploying, human controllers typically deviate from such normative procedures towards effective procedures (Dobbe, 2022). Therefore, such adaptation can be caused by for example time constraints, which can result in the human controller simply clicking the button (Dobbe, 2022). If the human controller simply clicks the button and does not consider broader implications, the same result as using CI/CD pipelines is reached, just adding an extra click.

To conclude, MLOps practice offers promising solutions to improve reproducibility, traceability, and accountability of ML systems. Furthermore, automatic monitoring of the ML system allows organisations to keep control over ML systems over time, and improve the ML system if needed. However, MLOps practice focuses solely on the technical functioning and making the ML lifecycle more efficient (Makarius et al., 2020). A deployed ML system does not operate in isolation, but is a subsystem of a larger sociotechnical ML system. This notion has to be an integral part of the entire ML lifecycle to be able to specify, develop and operate responsible and safe systems (Dobbe et al., 2021). Vulnerabilities in sociotechnical ML systems emerge from across the subsystems and interactions between those (Dobbe et al., 2021). Therefore, a technocentric view that could address technical or mathematical vulnerabilities in the ML system is not sufficient to secure the use of ML systems (Dobbe et al., 2021).The sociotechnical ML system view used in this master thesis research allows the researcher to identify vulnerabilities that could emerge in the development and

use of sociotechnical ML systems.

3.5 Conclusion Chapter 3

This section aims to answer the first sub-question:

1. What does the Machine Learning lifecycle look like in general, and what limitations can be identified from a sociotechnical ML systems perspective?

The ML lifecycle roughly consists of three stages, that all contain different activities:

- Experimental stage: contains the identification of the problem, collection of data, data analysis, feature engineering and data processing, building the model, and training, evaluating and validating the model.
- Deployment stage: contains the deployment of the model and integration of the model into its application
- Operations stage: contains monitoring of the model and its application

The ML lifecycle pictures a technical perspective of what it entails to develop an ML system, put it to production and operate it. To address the technical challenges that appear in the ML lifecycle, MLOps has emerged as a technical approach to automate, manage and speed up the ML model's lengthy operationalization by integrating DevOps practices into ML (Ruf et al., 2021).

However, the larger sociotechnical ML system the ML system becomes part of, is hardly considered in the ML lifecycle and MLOps. A too narrow technological view on the ML lifecycle may result in a problematic or incomplete specification of the system. This gives space for vulnerabilities to emerge in the sociotechnical ML system that may lead to harm.

The next chapter builds upon this knowledge by diving into a sociotechnical specification in the ML lifecycle and researching which vulnerabilities may emerge in the sociotechnical ML system that should be addressed in the specification.

Chapter 4

Specification and dimensions in sociotechnical ML systems

This chapter builds upon the knowledge gathered in 3, which presented an overview of the general ML lifecycle and MLOps practice, and their limitations. The chapter concluded that a narrow technical view can result in a problematic or incomplete specification of the system. This is a source for potential vulnerabilities to emerge in the resulting sociotechnical ML system. In this chapter, an approach to widen this narrow technological view is proposed by the introduction of sociotechnical specification in Section 4.1. Sociotechnical specification is meant to prevent, detect and solve vulnerabilities that emerge in the development and use of sociotechnical ML systems. Furthermore, to be able to guide practitioners in the sociotechnical specification, an understanding of the vulnerabilities that can emerge in sociotechnical ML systems must be reached. The vulnerabilities are identified by means of a integrative literature review, after which they are conceptualised in eight dimensions in which they emerge. The dimensions are the final theoretical conceptualisation, which is seen as vital to be addressed in a sociotechnical specification of sociotechnical ML systems. The dimensions are presented in Section 4.3 and the underlying vulnerabilities in Section 4.5. This chapter results in answering the second sub-question:

What sociotechnical dimensions should be addressed in sociotechnical specification of ML systems, based on potential vulnerabilities in sociotechnical ML systems?

4.1 Towards a sociotechnical specification in the ML lifecycle

As defined in the general ML lifecycle, the process starts with an identification of the problem, which entails defining the problem and why it is worth solving (Alla & Adari, 2021). This problem identification limits the view to the technical ML system and what it should solve, without considering the sociotechnical context the ML system becomes part of. At the same time, new forms of harm are generated by the integration of ML system in different domains, which are a product of the sociotechnical gap between what must be supported socially and what can be supported technically (Dobbe et al., 2021). An approach to bridge this gap is to widen the view and put more emphasis on the first step of the general ML lifecycle. To do so, guidance for a sociotechnical specification will be provided in this master thesis. A sociotechnical specification of the ML system is defined as the "proposed normativity of an ML system in terms of its featurization, optimization and integration, defining whom it is meant to serve, its purpose, and how it is to be evaluated and held accountable" (Dobbe et al., 2021, p.4). The sociotechnical specification thus specifies the following (Dobbe et al., 2021):

- featurization: the system's capacity to represent features of the environment in order to achieve a specified goal
- optimization: the designer's capacity to articulate how to more efficiently or appropriately complete a task
- integration: the capacity of stakeholders to oversee and incorporate the system's real-world performance

To develop ex ante ideas of what the role of sociotechnical specification could be and what sociotechnical specification entails, this section presents an overview of directions for sociotechnical specification.

4.1.1 Stakeholder involvement in sociotechnical specification

Sociotechnical specification emphasizes that developing ML systems for the use in social contexts means not just designing technology, but designing sociotechnical systems, which elements and interactions are all as central to the outcomes as the model outputs themselves (Green & Chen, 2019b). There is an emerging need for such specifications that are able to diagnose and resolve undesirable system performance, semantic ambiguity and political conflicts (Dobbe et al., 2021). All technological systems, and thus an ML system too, are political, as a collective deliberation of stakeholders is required to ensure its safety for everyone affected by it (Langdon, 1980). The start of the ML lifecycle requires an initial determination of the problem formulation and how it is to be tackled, feedback mechanisms to improve this initial determination and which stakeholders should be involved in the problem formulation (Dobbe et al., 2021). Stakeholders are all organisations and individuals who are directly or indirectly involved in, or affected by, ML systems. A subset of stakeholders is ML actors, who play an active role in the ML lifecycle (Dobbe et al., 2021). In sociotechnical specification, an understanding of the context of the envisioned ML system needs to be reached, including the positions of different stakeholders, their interests and how these relate to each other. Also, the impact of the ML system on social behaviour, broader societal implications, and positioning within existing legal frameworks needs to be understood (Dobbe et al., 2021). This requires active engagement of stakeholders, for which channels need to be created by which they can actively determine the ML system specification, rather than passively accept it (Unger, 1983). Additionally, getting stakeholders involved early in the project, fosters their confidence in the resulting ML system (Paleyes et al., 2020). In the sociotechnical specification, the roles and responsibilities should be determined across stakeholders (Dobbe et al., 2021). Currently however, most decisions in the development of ML systems are made by data scientists, while they should be based on an understanding of the application domain and the stakeholder needs (Vogelsang & Borg, 2019).

4.1.2 A program of requirements

In a sociotechnical specification, stakeholders should negotiate a program of requirements on both the process and outcomes (Dobbe et al., 2021). However, ML systems are characterised by a lack of specification of requirements (Kuwajima, Yasuoka, & Nakae, 2020). At the same time, most errors in operational software are caused by requirements flaws, incompleteness of requirements in particular (Leveson, 2012). Therefore, better techniques are needed to assist in determining requirements (Leveson, 2012). In ML research, often a good connection of the ML system to the domain of application is lacking, and appropriate determination of requirements as such (Dobbe, 2022). As systems are designed using a techno-centric approach, the complex relationships between the technical system and the people and processes it interacts with are not adequately considered (Baxter & Sommerville, 2011). To come up with a complete set of requirements, a system designer can identify vulnerabilities that could be present in the system, which can be translated into concrete requirements (Dobbe, 2022). The ultimate sociotechnical ML system designed should in turn be evaluated based on these requirements (Kuwajima et al., 2020).

4.1.3 Specification for emergence

The sociotechnical system of which a developed ML system becomes part, is not a pre-existing sociotechnical environment (Dobbe et al., 2021). Rather, the development of the ML system creates new situations that intervene on social life, which should be aware of when specifying this sociotechnical ML system (Dobbe et al., 2021). Values, such as safety and fairness, are emergent properties, that arise from the interactions among the system components (Dobbe et al., 2021). Therefore, it is vital to consider the whole sociotechnical ML system, including the ML system, stakeholders, decision-making processes and supporting infrastructure in the specification (Dobbe et al., 2021). This sociotechnical ML systems view can also explain how vulnerabilities in ML systems originate from the components in the sociotechnical system (Dobbe et al., 2021). Moreover, potential vulnerabilities emerge through the interaction between system components (Dobbe et al., 2021). To that extent, sociotechnical specification needs to be a cybernetic practice, in which feedback channels in the ML lifecycle should be established to refine the sociotechnical

specification to account for emergence (Dobbe et al., 2021). Because of the emergent dynamics within sociotechnical ML systems, validation of the ML model should be performed in its actual empirical context, rather than only applying the standard training and testing procedures (Dobbe, Dean, Gilbert, & Kohli, 2018). The validation of the ML system in the empirical context should look for emergent biases, overlooked specifications and other externalities (Dobbe et al., 2021). To safeguard the integration of the ML system in its sociotechnical context, forms of evaluation need to be established (Dobbe et al., 2021). The evaluation of ML systems needs to be expanded beyond evaluating the models themselves to investigating the whole sociotechnical ML system in which humans and ML system interact (Green & Chen, 2019b).

4.2 Theoretical framework building

As described in 4.1.3, vulnerabilities emerge in a sociotechnical ML system, which should be taken into account in the specification of these systems. However, there is no theoretical framework that covers these vulnerabilities in a constructive way, that can serve as scientific basis for the remainder of this research. Thus, this section presents the construction of a theoretical framework, based on a literature review on vulnerabilities that emerge in sociotechnical ML systems. Although theory building is not the ultimate goal in design science research, theory is needed as an intermediate artefact to build the design on(Winter, 2008).

To arrive at the theoretical framework, the Grounded Theory building method has been used. This method follows an inductive approach in order to generate or discover theory (Torraco, 2002). The theory evolves through continuous interplay between analysis and data collection (Torraco, 2002). As a result, Grounded Theory building allows new theoretical understandings to emerge from the data (Torraco, 2002). In this research, the data collected is data on vulnerabilities that emerge in sociotechnical ML systems identified in scientific literature. The analysis comprehends the interpretation of these vulnerabilities by combining them with knowledge of the ML lifecycle and sociotechnical specification. Finally, the resulted theoretical framework presents eight dimensions in which the vulnerabilities emerge, that should be addressed in the sociotechnical specification of sociotechnical ML systems.

4.3 Sociotechnical dimensions

As explained in 4.2, an inductive method is used to build a theoretical framework for this research. This results in the following theoretical framework, consisting of eights dimensions in sociotechnical ML systems in which vulnerabilities arise:

- Misspecification
- Machine error
- Interpretation
- Behaviour
- Adaptation
- Dynamic change
- Downstream impact
- Accountability

4.4 Mapping of sociotechnical dimensions in the ML decision-making process

This mapping illustrates the ML decision-making process within the sociotechnical ML system, in which the dimensions are placed at the process steps where vulnerabilities in these dimensions emerge.

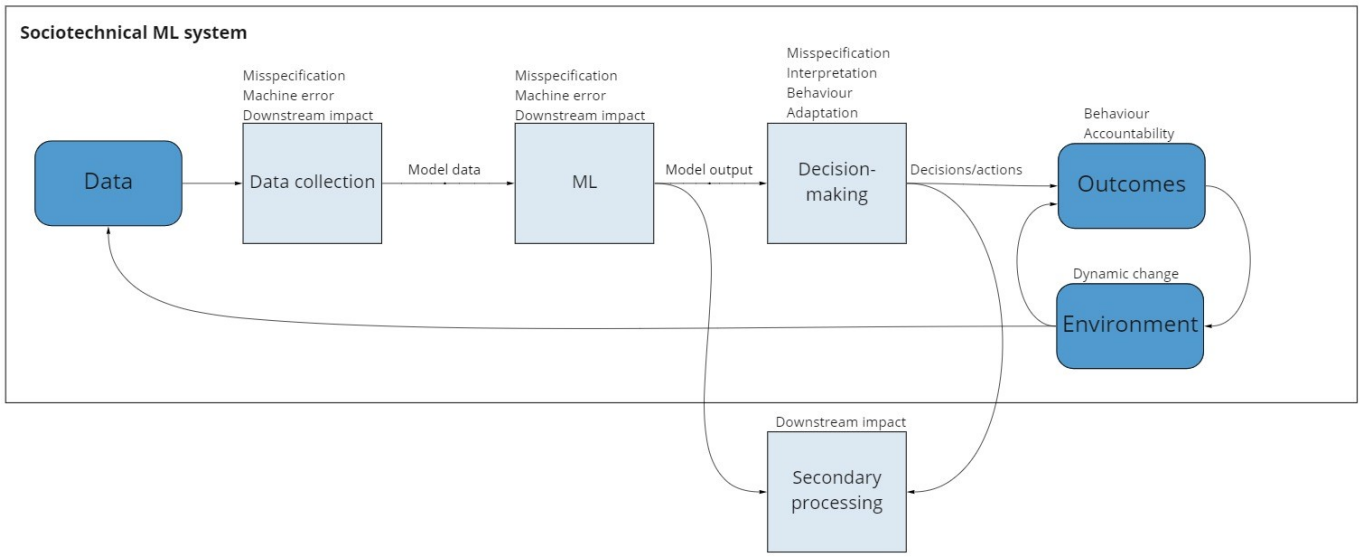

Figure 4.1: Dimensions mapping in the ML decision-making process

The following sections explain each of the sociotechnical dimensions.

4.4.1 Misspecification

Vulnerabilities can be caused by misspecification, which entails mistakes or gaps in the specification of ML systems. As ML systems are part of larger sociotechnical systems, this should be acknowledged in the specification of those systems. Facilitating the interaction between those components should be part of the specification, the absence of which could lead to ML systems that do not serve the needs of users and other stakeholders using the ML system. When relevant sociotechnical components such as people, processes and institutions are not sufficiently considered in the specification, this could lead to a sociotechnical ML system that has harmful consequences to individuals and organisations.

4.4.2 Machine error

Machine errors could occur in the ML system resulting in errors in the model output, and have impact on people and organisations. It is important to be aware of the types of machine error that could occur in the development of ML systems or in the ML model output and what the potential impact of errors could be to whom. Machine error can roughly be divided into two categories: machine incorrectness and machine bias. Machine incorrectness refers to a false negative or false positive output. Machine bias entails bias that is present in the model output, caused by for example a reproduction of past disparities or technical bias.

4.4.3 Interpretation

The interpretation of model output by a human decision-maker is a source of vulnerabilities. The model output can be misinterpreted due to a lack of understanding how the model output is generated. Further, humans can over-rely on the model output without the consideration of other factors, by which errors in the model are adopted in the final decision. Additionally, human decision-making brings about error that can be reduced or enforced by the introduction of an ML model in the decision-making process. Human error can be divided into two categories: bias and noise. Noise is an unwanted variability in professional judgement, whereas bias is the systematic error that is made by humans in the judgement of certain situations. ML systems can reduce noise and bias when designed and used properly, but can also confirm or even reinforce noise and bias. ML models can be hard to interpret for humans, due to their opaque nature. Different approaches to deal with the interpretability of models can infer vulnerabilities. Some models are so complex that it is impossible to understand for humans how decisions are made, which are often referred to as black-box models. An approach is to improve the interpretability of models is to make ML models explainable. The EU requires a 'right to explanation' of ML models (Rudin, 2019). However, it is not clear whether this explanation should be complete, accurate or faithful to the underlying model, making this an ambiguous requirement (Rudin, 2019). Lastly, effort could

be put in making ML models inherently interpretable. The way human decision-makers interpret model output can thus affect the ultimate decisions made using the ML system. This raises the question of how ML systems should be incorporated into decision-making processes (Green & Chen, 2019a).

4.4.4 Behaviour

When an ML system is integrated into a social context, the ML system will interact with a pre-existing social system, consisting of actors in that system (Selbst et al., 2019). To truly understand what the outcome of using an ML system is, it is vital to understand how this interaction takes place. The behaviour of actors can be affected by the introduction of an ML system, because using the ML system in the decision-making process can generate new incentives to behave in a certain way. Those actors could be human decision-makers using the ML system, or actors that are affected by the decisions made using an ML system. If human behaviour is delineated too much in the specification of the sociotechnical ML system, the risk emerges that if humans do not behave as expected, the sociotechnical ML system leads to unwanted outcomes. To prevent this, systems should accommodate the autonomy of users in contexts where this is wished.

4.4.5 Adaptation

In the specification of the system, it is not entirely possible to predict how users of the ML system will use the system in terms of rational work processes (Rasmussen, 2000). Developers of the ML system should be aware that operators or users of the system could deviate from the specified use of the system to address the complexity of the environment. A work environment in which an ML system is integrated is often complex, and humans that work in that environment have the ability to cope with this complexity by inventing clever strategies that do not match with what the system developers consider 'rational behaviour' of the user (Rasmussen, 2000). System developers should be aware that people can adapt in order to use a system. Therefore, assuming them to be rational and behave in a certain way is problematic.

4.4.6 Dynamic change

Once an ML system is put to production, Vulnerabilities could also emerge over time due to dynamic change. The world is continuously changing, in terms of data shifts, concept drifts, changes in regulations or organisational strategy. Dynamic change can have direct effect on the performance of an ML system and the output, for example by a decrease of accuracy. To address this decrease in performance, an ML model can be updated with new data and retrained. However, this can lead to a runaway feedback-loop that reinforce bias over time. Further, dynamic change might necessitate revisions in the sociotechnical ML system, for example to adhere to new regulation. To conclude, it is key to capture dynamic change in the sociotechnical ML system

4.4.7 Downstream impact

The introduction of ML systems in organisations, can have large downstream impact throughout the organisation and beyond. First, decisions made in one step in the ML lifecycle, could have impact on later steps in the ML lifecycle and the output of the ML system. Therefore, it is important for stakeholders to consider the downstream impact of decisions made in the ML lifecycle. Further, machine error in the model could work its way to secondary processes or systems in which the model output is used. Additionally, it could be human error in the final decision made using the model output that is used in secondary processes or systems. Moreover, if bias is inferred in the ML lifecycle, this could have downstream impact on the people that are ultimately affected by the decisions made using an ML system, or are affected by secondary processes or systems that are influenced by the ML system.

4.4.8 Accountability

Outputs of an ML system and decisions made using an ML system should be held accountable. However, it is often difficult to determine who carries the accountability of what part of the system. Undefined or unclear accountability causes vulnerabilities. Furthermore, accountability of systems could be lacking if model output and decisions are not reproducible.

4.5 Vulnerabilities in sociotechnical ML systems

Vulnerabilities are potential shortcomings in the ML system and sociotechnical ML system which may turn into hazards that cause harm to stakeholders. Hazards are defined as "[a sociotechnical] system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to an accident (loss)." (Leveson, 2012, p.184). Hazards are thus states that the sociotechnical ML system should never be in, and that have to be designed out of the system (Dobbe, 2022). To identify the vulnerabilities, an integrative literature review has been performed. The selected papers describe one or more vulnerabilities that can emerge in sociotechnical ML systems. To create a comprehensive overview, literature with a variety of research themes have been consulted. An overview of the selected literature and main theme can be found in Appendix A. The following paragraphs present an overview of the different vulnerabilities that can be present in sociotechnical ML systems.

4.5.1 Choosing ML as a solution

One should recognize that in not every situation, building an ML model is the solution to a problem. Firstly, if definitions of fairness are politically contested or shifting, it might not be possible to capture the facets of how it changes in the model (Selbst et al., 2019). Secondly, when there is not enough information about the social context, models are as likely to improve as to make the situation worse (Selbst et al., 2019). Therefore, it is essential to study what could happen when implementing the model, rather than implementing it just based on its potential to improve the situation (Selbst et al., 2019). It could also be that the attributes of a system are immeasurable, for example when those involve human psychology. In these situations, implementing a model could be not the right solution to a problem.

4.5.2 Framing of ML systems

Abstractions are essential to ML, but setting the abstraction boundaries too small, can result in unfair decisions (Selbst et al., 2019). One may define the system boundaries as an *algorithmic frame*, which entails the evaluation of the system on the algorithmic performance, for example the accuracy of the algorithm on training data. Widening the systems boundaries results in the *data frame*, which expands the frame to not just the algorithm but also the inputs and outputs of the final model (Selbst et al., 2019). While the *data frame* facilitates fairness considerations, this frame still eliminates the larger sociotechnical context of the ML system. Contrary, a *sociotechnical frame* does recognize explicitly that the ML system is part of the larger sociotechnical ML system and includes the decisions made by humans and institutions within the abstraction boundary. This expansion of system boundaries is essential to be able to evaluate fairness of sociotechnical ML systems. If this is not recognized and the ML system is evaluated as it is fully autonomous, while in reality it is part of a sociotechnical system with institutional structures and human decision-makers, harmful consequences may follow (Suresh & Guttag, 2021).

4.5.3 Function creep

Portability is usually pursued in ML, for example by reusing code to train algorithms or to provide "fair" ML systems by defining a definition of fairness that is portable (Selbst et al., 2019). However, portability causes that assumptions made about one sociotechnical context, are also used in other sociotechnical contexts. Since framing the system should entail the sociotechnical context of an ML system, assumptions should be made specific to this sociotechnical contexts. Therefore, a system designed for one sociotechnical context, is not portable between sociotechnical contexts (Selbst et al., 2019). If systems designed for one context are used in other contexts, this is called function creep, referring to a system or technology which function is expanded beyond its original specified purposes (Koops, 2021). Applied to ML systems, an example of function creep could be an ML system that has been specified and developed for one decision-making process, to be eventually being used for other objectives as well.

4.5.4 Mathematical fairness definitions eliminate societal nuances

Many efforts have been made in the Fair ML field to mathematically define fairness in order to incorporate fairness ideals into ML (Selbst et al., 2019). However, fairness in society is fundamentally vague, and limiting the notion of fairness to a mathematical formulation is problematic in

two ways. First, even if there would be an appropriate mathematical definition of fairness among potential definitions, it is impossible to determine this using purely mathematical means. The social context of the ML system determines what is fair and what is not. To illustrate, in one social context the consequence of a false positive prediction in an automated CV screening would mean a little extra work for the employer because a candidate is interviewed while he should be closed out in the CV screening. In the context of criminal justice, a false positive would mean that a prisoner would be held in prison longer, instead of being released. This illustrates that normative values determine what is fair in which social context (Selbst et al., 2019). The second problem is that fairness is a complex concept, for which no definition might be a valid way to describe it. Fairness may be procedural, contextual, and politically contestable, and mathematical definitions eliminate those nuances (Selbst et al., 2019).

4.5.5 Fail-safe mechanisms and plan B procedures

Like all technologies, ML systems could sometimes fail in their functioning (Dobbe, 2022). Therefore, solely relying on the ML model's safety constraints is dangerous (Dobbe, 2022). Therefore, there should be taken safety measures to compensate an ML system's malfunctioning, such as fail-safe feature to kick in when the ML system fails to prevent harm (Dobbe, 2022).

4.5.6 Wrong assumptions about operations

When an ML system is developed, assumptions about the system in operation are made. However, it can be that the assumptions made are not appropriate (Leveson, 2012). It could also be that the sociotechnical context changes over time, causing initially correct assumptions to become incorrect. (Leveson, 2012).

4.5.7 Setting the threshold for a positive vs negative output

The use of ML models can result in false positive or false negative output (Janssen, Hartog, Matheus, Yi Ding, & Kuk, 2020). False positive output (or Type I mistake) means that the model output indicates that a given condition is present, while it is actually absent. A false negative output (or Type II mistake) means that the model output indicates that a given condition is absent, while it is actually present (Janssen et al., 2020). ML models can be tweaked to increase or decrease the number of false positives or false negatives. This involves trade-offs from a societal point of view about what is desirable. For example, in the case of searching for a criminal, many false positives mean that innocent people are classified as criminals, but decreasing the number of false positives can in turn increase the number of false negatives, which means that criminals might not be detected as criminals (Janssen et al., 2020). Setting the threshold for a positive versus negative outcome could have different impact for different groups. This could impose fairness issues that were not detected before deployment (Veale & Binns, 2017).

4.5.8 Reproduction of past disparities

In supervised learning ML models, labelled data from previous decision-making is used to train the model (Veale & Binns, 2017). ML models are supposed to discriminate between data points, but some logics of discrimination are not socially acceptable. Thus, if the historical data that the ML model is trained on reflects unwanted discrimination, it is likely that these patterns will also be shown in the model's predictions. Therefore, there is a risk of past disparities to be reproduced using an ML model (Veale & Binns, 2017). For example, the widely-used COMPAS risk assessment tool wrongly labelled black defendants as future criminals twice as much as white defendants, due to historical discrimination embedded in the training data (Green & Chen, 2019a).

4.5.9 Technical bias

Besides the aforementioned bias in the historical data, other sources of bias occur when developing and employing ML systems (Dobbe et al., 2018). Technical bias is bias that is caused throughout the development stage of turning data into a model that can make predictions. Choices made in this stage can infer bias, for example when setting the scale of model variables (e.g. ordinal or nominal), choosing the type of model that is used, or when the model is optimized to certain objectives (Dobbe et al., 2018). Making these choices requires justification and often value judgement, which are context-specific and often ethical. Overlooking these questions is practice can have harmful consequences, especially in high-stakes domains (Dobbe et al., 2018).

4.5.10 The detection of incorrect output of black box models

ML models can be black boxes, that do not explain their predictions in a by humans understandable way (Rudin, 2019). These models' learned rules are so complex and non-linear that they are practically inexplicable, even to the model developers themselves, let alone to end-users or policymakers (Zejnilović et al., 2021). The use of black box models in high-stakes domains can have severe consequences, because incorrect decisions based on an inexplicable model are hard to detect.

4.5.11 Explainers can be inaccurate

Recently, there is a lot of work on explainable ML, where a second (post hoc) model is developed to explain the first black box model (Rudin, 2019). However, since a second model is developed, this model could be inaccurate as well. If an explainable model is correct 90% of the time, one cannot know whether an explanation is correct and whether to trust the explanation or the original model (Rudin, 2019).

4.5.12 Explainers do not consider the context

Also, current research in ML explainability is mainly focused on technical issues, such as developing measures to explain models and their outputs, and may not adequately consider the variety and complexity of the contexts where the ML systems are deployed (Zejnilović et al., 2021). Explainability tools may not necessarily provide the expected outcome when used in a real-life setting (Zejnilović et al., 2021). Therefore, ML system developers deploying explainers should consider the real-life setting including individual, technical, institutional, and political factors users are coping with. If not, explainers might lead to outcomes that deviate from the expectations (Zejnilović et al., 2021).

4.5.13 An explanation is not a general attribute

Moreover, different communities understand explanation in substantially different ways (Kohli, Barreto, & Kroll, 2018). To an ML researcher, an explanation is a description of the operation of the model, which covers the mechanisms used to relate inputs to output (Kohli et al., 2018). In the social science, there is a robust notion of how explanations should behave. Explanations should be causal, explaining why an output was reached or an event occurred; they should be contrastive, explaining why event X happened over event Y; they should be selected, which means they should be based on a few key causes rather than complete descriptions of a mechanism; and they should be social, which means they are meant to transfer knowledge about the system they are explaining (Miller, 2019; Kohli et al., 2018, p). Therefore, explanations should be contextual, and are not just a presentation of associations and causes. While an output may have many causes, the person that requires an explanation often cares only about a small subset, relevant to the context (Miller, 2019). Therefore, ML models cannot be explained at a general level, but an explanation should be suitable for the stakeholder that requires an explanation.

4.5.14 Inherently interpretable ML models are challenging

Another way to prevent harm by black box models is to develop inherently interpretable models, instead of trying to explain black box models (Rudin, 2019). Interpretability is domain-specific, so it has no single definition. There is a spectrum of interpretability between fully transparent models where one can understand how all the variables are jointly related to each other, and models constrained in model form, so that it is either useful to someone, or obeys structural knowledge of the domain (Rudin, 2019). For example, models that are forced to increase as one of the variables increases or models that prefer variables that domain experts have identified as important (Rudin, 2019). There is a widespread belief that more complex models are more accurate, so that complicated black box models are necessary for top predictive performance (Rudin, 2019). However, if the data are structured, with good representation in terms of naturally meaningful features, this is often not the case (Rudin, 2019). In these cases, there is often no significant difference in performance between more complex models (e.g. deep neural networks, boosted decision trees, random forests) and much simpler models (e.g. logistic regression and decision lists) after pre-processing (Rudin, 2019). However, there are currently multiple challenges that prevent practitioners to develop inherently interpretable ML models. Firstly, companies make profits from the intellectual property that is created by a black-box model, as they charge per

prediction (Rudin, 2019). Secondly, to develop interpretable models, significant effort is needed in terms of computation and domain expertise. Lastly, to uncover 'hidden patterns', which is often called in favour of black-box models, an ML researcher has to be able to both create accurate and interpretable models, which is a difficult optimization challenge (Rudin, 2019).

4.5.15 Deployment bias

In DSSs, human decision-makers make the final decision, supported by the prediction made by the ML system. Even though the ML model's output could be unbiased according to certain metrics, the human decision that follows could still lead to biased and thus unfair decisions (Balayn & Gürses, 2021). Deployment bias arises when users introduce unexpected behaviour that affects the final decision (Suresh & Guttag, 2021). One example is confirmation bias, which means that human decision-makers seek for evidence or interpret evidence in ways that are in line with their existing beliefs or expectations (Nickerson, 1998). In other cases, people charged with using DSSs ignore or resist the model output (Green & Chen, 2019b). Lastly, the introduction of DSSs can prompt people to alter their behaviour, as they may overly rely on the ML system's output or focus on different goals due to incentives that are created by the introduction of the system (Green & Chen, 2019b). For example, they follow the ML system without considering contradictory information, leading to decisions that are not based on an analysis of all available information, but biased towards the model output (Green & Chen, 2019a; Parasuraman & Manzey, 2010). To incorporate the ML system's predictions, a user interface has to be developed which should be used the human decision-maker (Suresh & Guttag, 2021). Most user interfaces involve simply presenting the model output to a human decision-maker, relying on the person to interpret and incorporate that information (Green & Chen, 2019a). It is challenging to prevent deployment bias, but designing ML systems that help users balance their faith in model predictions with other information and judgements is an important part. This could involve choosing a model that is human interpretable or developing interfaces that help users to understand model uncertainty and how predictions are to be used (Suresh & Guttag, 2021).

4.5.16 Noise in professional judgement

Besides consequences due to human bias, judgements by human decision-makers can undesirably vary from one individual to the next, which is called noise (Kahneman, Rosenfield, Gandhi, & Blaser, 2016). Noise can even occur in the judgement of one individual from occasion to occasion, for example caused by irrelevant factors as the weather or mood (Kahneman et al., 2016).

4.5.17 Unclear responsibilities of decisions

Reliance on ML systems could change people's relationship to the decision-making task, by creating a 'moral buffer' between their decisions and the impact of those decisions (Green & Chen, 2019a). This can lead to human decision-makers to let go of a sense of responsibility and subsequently accountability, because they have the perception the ML system is in charge (Green & Chen, 2019a). Besides that, data scientists often express that they do not bear responsibility for the social impact of their models. Those phenomena can result in situations where both the data scientists developing the ML models and the users of the ML models think the other to be primarily responsible for the outcomes (Green & Chen, 2019a). This scenario should be avoided.

4.5.18 Data shifts

Data shift are changes in the input data distribution of the model. In the experimental stage, ML models are trained and tested on training and test data. But when the ML model is used in operation, the operational data that is fed to the model as input data tend to change over time (Kuwajima et al., 2020). This phenomenon is called data shifts, which might lead to decreased accuracy and fairness of the ML system over time (Balayn & Gürses, 2021). Data shifts may arise due to several reasons. The populations on which the models are applied might change over time or the way data is captured differs between training and operation, making the data input different from the training data (Balayn & Gürses, 2021). Also, changes in the real-world lead to changes in the data distributions that represent real-world concepts (Tsymbal, 2004). This way, the input data distribution changes, which might lead to model behaviour changing in unwanted ways (Brom, 2021). This would require data shift detection.

4.5.19 Concept drift

A related vulnerability is concept drift, which are changes in how well the model understands the relationship between input and output. Often, the cause of this change is hidden, and not known upfront, for example change in ways humans think and behave, which leads to changes in what the model is expected to infer (Balayn & Gürses, 2021).This would require concept drift detection, sets of techniques that can be used to automatically detect shifts in distributions potentially relevant to the model's task (Veale, Kleek, & Binns, 2018).

4.5.20 External factors change over time

If changes in an organisation or society, for example changes in rules and regulations are established over time, this brings challenges for the development of ML systems (Veale et al., 2018). Awareness of these changes and adequate communication and preparation for them are essential but far from straightforward (Veale et al., 2018).

4.5.21 Runaway feedback-loops

Besides pre-existing bias in data and technical bias occurring in the development of ML systems, bias can also arise when the ML system is altered using feedback from its use (Dobbe et al., 2018). For example, in predictive policing, where discovered crime data (e.g. the number of arrests) are used to predict in which areas new crimes will arise, the police surveillance can be intensified in those areas (Dobbe et al., 2018). This way, it is likely that the discovered crime rate increases. If this new discovered crime rate is used for updating the model, the police will be repeatedly sent back to that area, regardless of the true crime rate (Dobbe et al., 2018). This shows that feedback used to update the model can impose emergent bias over time, which is called a runaway feedback-loop.

4.5.22 Lack of reproducibility

Developing and operating ML systems consists of many steps, often performed by multiple people. If the ultimate decisions that are made based on an ML system are not reproducible, it is not possible to trace back the cause when something goes wrong (Ruf et al., 2021). To reproduce output, training data should be versioned, experiments should be versioned, models should be versioned, it should be tracked which models is used for which prediction, and which input data is used for which prediction (Ruf et al., 2021; Kohli et al., 2018, p).

4.5.23 Decentralized data collection

When models are developed in different parts of an organisation than data collection efforts, it can easily happen that changes in data collection practices occur, without the people responsible for model performance being aware of those changes (Veale et al., 2018). It can be that data collectors not even know that the data is used for an ML system, particularly when data is collected in a decentralized manner, for example by auditors or police patrol officers (Veale et al., 2018). Better communication between data collectors and ML developers might help, but becomes increasingly difficult if more and more models are developed. Moreover, changes in data collection might be not explicit at all, but can emerge from cultural change or day-to-day choices (Veale et al., 2018). As one can see, it might be impossible to communicate all changes in the data collection. Another approach could be to detect changes in the data distribution itself, by means of concept drift detection (Veale et al., 2018).

4.5.24 Data cascades

Data cascades are compounding events that cause negative, downstream effects from data issues, that result in technical debt over time (Sambasivan et al., 2021). Data cascades are triggered when ML practices undervalue the importance of data quality, while data largely determines performance, fairness, robustness, safety and scalability of ML systems (Sambasivan et al., 2021). If not enough effort is put in ensuring data quality in the beginning, this will cause negative impacts in the remainder of the ML lifecycle. Data cascades are typically triggered in the beginning of the ML lifecycle, but appear unexpectedly when models are deployed and used in production, resulting in harm to people, discarded models, and redoing data collection (Sambasivan et al., 2021).

4.6 Conclusion Chapter 4

This section aims to answer the second sub-question:

What sociotechnical dimensions should be addressed in sociotechnical specification of ML systems, based on potential vulnerabilities in sociotechnical ML systems?

The sociotechnical dimensions that should be addressed in sociotechnical specification of ML systems are:

- Misspecification
- Machine error
- Interpretation
- Behaviour
- Adaptation
- Dynamic change
- Downstream impact
- Accountability

These dimensions were specified based on 24 vulnerabilities that have been identified in selected literature. The dimension capture all vulnerabilities, but do not have a one-on-one relation. Rather, vulnerabilities may arise in multiple dimensions, or a combination of those. The dimensions serve as a scientific knowledge base to the next phase of the research, in which interviews are performed in the application environment. In the next chapter, Chapter 5, the use cases selected for this research within the application environment are introduced. Thereafter, an analysis of the dimensions within the use cases is presented in Chapter 6.

Chapter 5

Use case descriptions and specification in practice

As described in the research approach, the second phase of the research entails the identification of opportunities and problems in an actual application environment (Hevner, 2007). These opportunities and problems in turn can serve as input for the design cycle (Hevner, 2007). The application environment for this research is the use of ML in the financial domain, as explained in Chapter 1. In order to gather data about the application environment, two use cases within this application environment are studied. This chapter presents how the use cases are selected, how the interviewees for this research are selected, and a description of both use cases in Section 5.1 up to Section 5.5. Section 5.6 presents the results of the analysis of the specification process in the use cases. This will lead to answering the third sub-question:

What does specification in the ML lifecycle look like in practice, based on two use cases in the financial sector?

5.1 Selection of use cases

The second phase of the research aims to create an in-depth understanding of the practice of specification of ML systems and the addressing of the sociotechnical dimensions identified in Chapter 4. To do so, the financial domain has been chosen as the application environment to dive into. There is great variety within the financial domain in organisations that use ML, the maturity of the organisations regarding ML and the context in which ML models are used. To get insight in the differences and similarities, two ML use cases with different characteristics have been selected for analysis. A use case is a specific situation in which ML is developed and used. The main characteristics of the selected use cases are summarized in the table below.

Table 5.1: Overview use case characteristics

	Use case: Financial crime detection	**Use case: Email marketing**
Bank	Bank A	Bank B
Objective ML system	Reduce the false-positive rate of alerts and find new types of alerts on financial crime	Personalising marketing in order to increase the conversion rate towards investing products
Type of decision-making process	Decision support system	Automated decision-making with human controller
Outsourcing partners involved	None	ML engineering/consulting firm

In the research, the development of the ML system as well as the use of the ML system in practice are studied. The use cases researched are situated within different banks, serve a different goal, the type of decision-making process varies and the ML models were either developed completely in house or by partnering with and external firm. To gain a rich understanding of the use cases, a variety of stakeholders involved are interviewed.

5.2 Selection of interviewees

For this research, eighteen professionals have been interviewed by means of semi-structured interviews. Eleven professionals directly or indirectly involved in the use cases within the banks and the external ML engineering/consulting firm have been interviewed to gather insights into the application environment. The interviewees were selected as follows. The stakeholders identified in Section 3.1.5 are used as a first step to select stakeholders to interview. Also, the sociotechnical ML systems perspective that is taken throughout this research provided directions for the selection of the interviewees. The research focuses on the specification and development, as well as the integration and the use of an ML system in its sociotechnical context. Taking this into account, interviewees involved in the specification of the ML system, the development of the ML system, the use of the model in a decision-making process, and the compliance are selected. As the stakeholders in practice sometimes fulfil more than one role presented in Section 3.1.5, the following overview presents the interviewees and their roles as defined in 3.1.5. Two interviewees within bank A, I6 and I7, are not directly involved in the development of this use case, but provided overarching insights on integration of use cases and responsible AI within the bank in general. Further, in the email marketing use case, no compliance officer has been interviewed, due to in-availability. The names of the interviewees and the names of the organisations they work at are not disclosed due to privacy reasons.

Besides stakeholders involved in the practical use cases, seven interviews have been performed with stakeholders not directly involved in the use cases and outside the banks. These interviews were performed to get a broader understanding of the implications of ML systems related to the sociotechnical dimensions identified on the customers of the banks and the larger public. Representatives of several civil society organisations and regulatory bodies have been interviewed to this extent. These organisations are Waag, Amnesty International, Privacy First, Platform Bescherming Burgerrechten, Bits of Freedom, The Dutch Data Protection Authority (Autoriteit Persoonsgegevens) and De Nederlandsche Bank (DNB). The names of the representatives are not disclosed due to privacy reasons. Table 5.2 presents an overview of all interviewees, including their function, stakeholder role(s), use cases involved, organisation and interview ID.

Table 5.2: Interviewee information

Interviewee function	Stakeholder role(s)	Use Case	Organisation	Interviewee ID
Manager data science	Product owner	Financial Crime Detection	Bank A	I1
Lead data scientist	Data scientist	Financial Crime Detection	Bank A	I2
Transaction monitoring analyst	End-user, subject matter expert	Financial Crime Detection	Bank A	I3
Data Engineer	Data engineer	Financial Crime Detection	Bank A	I4
Privacy Officer	Compliance officer	Financial Crime Detection	Bank A	I5
ML engineer	ML engineer, MLOps engineer	Use case overarching	Bank A	I6
Enterprise advisor data science	Responsible AI integration	Use case overarching	Bank A	I7
Manager marketing intelligence	Product owner	Email marketing	Bank B	I8
Marketing intelligence Analyst	End-user, subject matter expert	Email marketing	Bank B	I9
Marketeer	subject matter expert	Email marketing	Bank B	I10
ML engineer/project leader	Data scientist, ML engineer, MLOps engineer	Email marketing	External ML engineering/consulting firm	I11
Representative	Reflection of technology	-	Waag	I12
Representative	Addressing human rights	-	Amnesty International	I13
Representative	Addressing privacy protection	-	Privacy First	I14
Representative	Addressing legal protection on civil rights	-	Platform Bescherming Burgerrechten	I15
Representative	Addressing human rights	-	Bits of Freedom	I16
Representative	Regulatory body	-	Autoriteit Persoonsgegevens	I17
Representative	Regulatory body	-	DNB	I18

The interview protocols used in this research are presented in Appendix B and Appendix C

5.3 Use case description: Financial crime detection

Banks are obliged by law (Wet ter voorkoming van witwassen en financiëren van terrorisme (Wwft) to report unusual transactions to the Financial Intelligence Unit Nederland (Financial Intelligence Unit, n.d.). Also, banks have to do customer investigations (AFM, n.d.). As such, banks have a responsibility to detect suspicious behaviour in order to mitigate the risk of money laundering and financing terrorism. At bank A, there was a rule-based system in place that sent out alerts on suspicious transactions to the transaction monitoring analysts within the bank. A transaction monitoring analysts then start an investigation of the customer(s) involved. The transaction monitoring analysts assesses whether the customer is indeed is performing suspicious activities, after which the customer is further reviewed or not by another department within the bank. Finally, the bank reports such customers to the Financial Intelligence Unit. However, the transaction monitoring analysts encountered the problem that the rule-based system provided many alerts that were in the end not assessed as suspicious activity. In other words, the rule-based system knows many false-positive alerts. As the bank wants to detect as many suspicious activities as possible, but in

an effective and efficient manner, this was problematic.

To improve this process, there were two ML models developed. The first ML model was developed to decrease the number of false-positives from the rule-based system, by predicting whether an alert is actually a true-positive alert or a false-positive alert. The predicted true-positive alerts are then pushed forward to the transaction monitoring analysts to investigate. The transaction monitoring analysts investigate the alert to detect potential financial crime. As such, they ultimately assess whether the alerts are actual true-positive alerts or not. The second model is a model that detects new suspicious behaviour, that is not detected by the rule-based system, to be pushed forward to the transaction monitoring analysts to investigate. The transaction monitoring analysts use the models' output as a source of information for their investigations.

The interviewees for this use case are:

- I1: Manager of the data science team that develops ML models for the detection of financial crime within Bank A. Sets priorities for new use cases and aligns this with the business. Carries the delegated execution of model ownership from higher management.
- I2: One of the lead data scientists within the data science team. Technical lead of the two ML models within the use case. Responsible for keeping the models running in operation and the development of new model versions.
- I3: Transaction monitoring analyst is end-user of the models' output and has been involved in the operational implementation of the model into the decision-making process of the transaction monitoring analysts
- I4: Data Engineer has migrated all features used in the ML models within the use case to a central feature store within the bank A, which allows the features to exist independently from the models.
- I5: Privacy officer within bank A. Assesses privacy risks and compliance of new ML use cases and model versions.
- I6: ML engineer supports data scientist by bringing ML models to production and keeping them in operation by providing monitoring, AB testing, and explainability capabilities. Besides that, I6 focuses on Responsible Data Science and governance of ML within Bank A in general.
- I7: Enterprise advisor data science leads the responsible AI activities within bank A. Facilitates the development of use cases from beginning to end by a model management framework.

5.4 Use case description: Email marketing

At bank B, there was a wish to better inform private customers and in a more personalized way about investing, in order to ultimately increase the conversion rate from customers without an investing account towards investing. Before an ML system was developed, the bank did a marketing campaign for investing a few times a year, in which every private customer without an investing account received an email with information about investing. To increase the conversion rate, the bank wanted to move towards a more year-round way of marketing for investing. In order to do so, the Marketing Intelligence team of the bank has partnered with an external ML engineering/-consulting firm to develop an ML system to facilitate this. The ML model developed predicts for every relevant customer at a certain moment what the probability of conversion to investing is. If the probability is higher than a set threshold, this customer will be presented in the model output. The Marketing Intelligence Analyst received this output and pushes the customers to a emailing system, from which these customers are emailed. The Marketing Intelligence Analyst performs some additional checks before the emailed are being sent.

The interviewees for this use case are:

- I8: Manager Marketing Intelligence is responsible for the Marketing Intelligence team's analysis, which includes the ML system.

- I9 Marketing Intelligence analyst was involved in the specification of the use case and functions as a human-in-the-loop to check model output and push the selected customers to the emailing program
- I10 Marketeer is subject matter expert on investment marketing and was with this expertise involved in the use case development
- I11 The external ML engineer/project manager developed the ML model and put it into production within the banks infrastructure.

5.5 Civil society organisations and regulatory bodies

The following representatives of civil society organisations have been interviewed.

- I12 Representative of Waag focuses on the reflection on technology, among which ML. The organisation has a critical constructive role and has knowledge of technology in-house.
- I13 Representative of Amnesty International focuses on policies and regulations, primarily in governments and researches which risks for human rights the use of ML systems has
- I14 Representative of Privacy First has a focus on the protection of privacy and addresses potential privacy violating regulations or outcomes of regulations
- I15 Representative of Platform Bescherming Burgerrechten mainly focuses on how data-driven work by the government affects people's legal protection and has a critical attitude towards it.
- I16 Representative of Bits of Freedom focuses on the impact of AI on human rights, by influencing policy on both national and European level.

The interview data gathered from the civil society organisations was not always specific for banks or the financial sector. The organisations are not primarily focused on the financial sector, but provided a lot of insight into the perspective of these organisations on the use of ML in relation to citizens in the Netherlands. This way, the voice of the citizens about whom decisions are made using ML systems is represented. The following representatives of regulatory bodies have been interviewed:

- Representative of Autoriteit Persoonsgegevens stimulates compliance of the GDPR within organisations and their systems. The Autoriteit Persoonsgegevens regulates the GDPR, which applies to ML systems that process personal data.
- Representative of DNB focuses on establishing AI supervision in the financial domain. The role of DNB as such is to identify the risks of using ML and the impact on the financial domain.

5.6 Specification in the ML lifecycle in practice

In Section 4.1, the argument is made that in order to develop and operate ML systems that will be used in an application domain, not only the ML systems, but the larger sociotechnical ML system has to be designed. In order to do so, sociotechnical specification is at the core. In the following paragraphs, an analysis of both use cases will follow to get an understanding of how specification in the ML lifecycle currently looks like. These insights from the application environment will serve as input to the design cycle.

5.6.1 Specification in the financial crime detection use case

Problem identification
The financial crime detection use case started with a problem identified. The problem identified was that the rule-based system that generates alerts with suspicious behaviour had a lot of false-positives output. This made the transaction monitoring analysts to investigate alerts that do not actually contain suspicious behaviour (Personal Communication Lead Data Scientist, January 12, 2022). Investigating alerts unnecessarily results in a waste of time and money. To identify and

define this problem, the data scientists talk to experts in the application domain (Personal Communication Manager Data Science, January 12, 2022).

Program of requirements
Before the development of a model starts, the involved data scientists define a program of requirements together with other stakeholders within the bank (Personal Communication Manager Data Science, January 12, 2022). Most importantly, models should provide business value, for example by decreasing the workload of the transaction monitoring analysts or by detecting new forms of financial crime (Personal Communication Manager Data Science, January 12, 2022). There were no specific requirements for the performance of the models defined upfront (Personal Communication Lead Data Scientist, January 12, 2022). Besides the requirements that are specified by the data scientists, the bank has many standard requirements that need to be adhered to when a new ML system is developed. To guide the data scientists throughout the ML lifecycle, there is a bank-wide framework in place that needs to be followed. This framework describes at which point they should apply for different approvals from for example legal, privacy, risk, and ethics (Personal Communication Lead Data Scientist, January 12, 2022). For example, there is an initial privacy check that needs to be performed before model development starts (Personal Communication Privacy Officer, January 17, 2022). All necessary approvals have to be given before the model is put to production.

Although the data scientists defined a program of requirements together with stakeholders, the transaction monitoring analysts, who are the end-users of the ML system, were not involved from the beginning. This led to a specification of the feature descriptions that was very technical, and very difficult to understand for the transaction monitoring analysts (Personal Communication Transaction Monitoring Analyst, January 14, 2022). As such, the transaction monitoring analysts had to make a translation file to make the descriptions understandable. If a transaction monitoring analyst had been involved from the beginning of the use case development, the feature descriptions could have specified differently, to be direct understandable for the analysts (Personal Communication Transaction Monitoring Analyst, January 14, 2022).

Model validation
To validate a model before it is put to production, there is an independent model validation team that validates the ML model, checks for errors and evaluates choices made (Personal Communication Lead Data Scientist, January 12, 2022). What is notable, is that the validation is aimed at the ML model itself, and not at the ML system in its application domain. The transaction monitoring analyst noted that they are currently working with a new model that has not been tested by the actual end-user in the case management system the transaction monitoring analysts use to handle the alerts. This has resulted in model outputs that contains a lot of errors, which the transaction monitoring analysts have to work with. Therefore, the model should be tested in the actual case management system, by the transaction monitoring analysts being the end-users (Personal Communication Transaction Monitoring Analyst, January 14, 2022).

Monitoring and evaluation of the sociotechnical ML system
There are monitoring and evaluations in place for the ML system. The ML models are run monthly, and before the output is directed to the case management system, there is a performance monitoring in place (Personal Communication Lead Data Scientist, January 12, 2022). In this performance monitoring, it is checked whether the model performs comparable to the training phase. Furthermore, the models have a certain period of validity, after which either extensive checks on the performance have to be performed, or a new model version has to be delivered (Personal Communication Lead Data Scientist, January 12, 2022). Besides the performance, the risks that have been identified by risk officers are also automatically monitored and evaluated over time, because the risks can change over time (Personal Communication Privacy Officer, January 17, 2022). However, there is no evaluation in place for the interaction between the ML system and the end-user (Personal Communication Manager Data Science, January 12, 2022). For the ultimate investigation reports by the transaction monitoring analysts, there is again evaluation in place. Up to three reports are checked every two weeks for every analyst, to check whether they have identified the present risks and took the right follow-up steps (Personal Communication Transaction Monitoring Analyst, January 14, 2022). As such, the distinct components within the sociotechnical ML system are evaluated, but the sociotechnical ML system containing the interactions is not.

Conclusion of specification in the financial crime detection use case

There are several elements of a sociotechnical specification present in this use case. First, the use case knows a clear problem identification as starting point. Besides that, requirements are defined, and the data scientists are guided with many requirements that are bank-wide defined. Further, an independent model validation team checked for errors and choices made concerning the models before they were put to production. Lastly, there are extensive monitoring and evaluation tools and procedures in place to safeguard the models over time.

Despite the identified elements presented in the use case, several shortcomings in the sociotechnical specification have been identified as well. First, the end-users of the ML system were not involved in the specification of the sociotechnical ML system. Second, the description of features were specified by the data scientists, which resulted in descriptions that the end-users could not understand, while they need them when using the ML system's output. Third, there are many approvals needed and processes to follow to safeguard the integration of new ML systems within bank A, but there is no guidance on how to design the decision-making process between model output and final decision (Personal Communication ML engineer, January 24, 2022). Ultimately, an ML system could adhere to all requirements, but if the decision-making process it becomes part of is not adequately specified and evaluated, this could still lead to poor outcomes (Personal Communication ML engineer, January 24, 2022).

5.6.2 Specification in the email marketing use case

Problem identification
The email marketing use case started with a clear initial problem identification. As the bank loses money on payment en saving accounts, the bank saw to urge to increase the customer base within investing. The specification of the use case was fairly organic. There was not a very specific goal to develop an ML model that needs to meet certain criteria at first, but the goal was to increase investment customers (Personal Communication, Marketeer, January 14, 2022). For the existing customer base, the bank does a general marketing campaign for investing a few times a year to inform them and activate them to start investing. Eventually, they figured that if the marketing would become year-round by targeting the most promising customers every week, the conversion rate to investing could increase. To determine which customers are the most promising to convert, ML came in.

Stakeholder involvement
In the specification of the use case, a few stakeholders were directly involved. The marketing intelligence manager, marketing intelligence analyst, marketeer and ML engineer/project manager from an external ML engineering/consulting firm. Those stakeholders have all a different expertise and perspective, and were actively engaged during the project. Before the development of the use case actually started, the external engineering/consulting firm started with a potentiality analysis to determine whether developing an ML model would be valuable. To do so, they interviewed the above-mentioned stakeholders and developed a business case with the collected information to conclude with the expected commercial results of the use case, including metrics to measure these. Further, they investigated whether the required data are available, how an ML system could be integrated in the technical infrastructure and how much time was needed to develop the ML system (Personal Communication external ML engineer/project manager, January 13, 2022).

Program of requirements
After the use case was concluded to be worth developing, a program of requirements was negotiated with the involved stakeholders. The requirements were commercial, technical, organisational and compliance oriented. Every stakeholder provided requirements from their own expertise. The scope of the specification of requirements was mainly ML system and commercial results oriented.

Feedback channels
During the development of the ML system by the ML engineering/consulting firm, there were feedback moments in which the variables, parameters, the threshold for the output were discussed to fine-tune and tweak the model (Personal Communication marketing intelligence analyst, January 21, 2022). The feedback moments thus not particularly reconsidered the sociotechnical specification as suggested by Dobbe et al. (2021).

Specification for emergence
As described in Section 4.1.3, a developed ML system becomes part of a sociotechnical system that

is not pre-existing, but emerges through interactions between system components. To prevent, detect and resolve vulnerabilities that can emerge over time, it is vital to take this emergence into account in the specification. Looking at this in the email marketing use case, a good approach is that validation of the ML system is performed in its actual empirical context. After the ML system was developed, the external ML engineering/consulting firm did a validation period of running the model in the background in the application domain, to see whether it satisfied the requirements and expectations, before actually operationalizing the ML system (Personal Communication external ML engineer/project manager, January 13, 2022). This validation was very ML system oriented, and did not take the interaction with human operators and other human agents into account (Dobbe et al., 2021). Moreover, emergent biases, overlooked specifications or other externalities were not explicitly mentioned by the interviews as being part of the validation. Besides that, the sociotechnical ML system that emerges involves a human operator that receives the model output, performs a few basic checks, and pushes the customers to be emailed to an email program. It was not a conscious decision to include the human operator. Rather, the human operator was introduced along the way because the technical systems could not interact (Personal Communication marketing intelligence manager, January 11, 2022).

Evaluation of the sociotechnical ML system
To safeguard the integration and use of the ML system in its sociotechnical context, forms of evaluation of the whole sociotechnical ML system should be specified (Dobbe et al., 2021). In the email marketing use case, there have been specified several forms of evaluation. In the first period after the ML system was put to production, the external ML engineering/consulting firm intensively evaluated the results of the model. For example, the model predicted that a relatively high number of customers had to be emailed at first. To keep this number manageable, the threshold on the model output on to be emailed/not to be emailed was adjusted (Personal Communication Marketeer, January 14, 2021). After that first period, there are periodic evaluations held a few times per year with the direct stakeholders. This evaluation is a commercial evaluation on the metrics defined in the business case at the start of the use case project. These metrics are for example the conversion rate to investing and the amount of money that customers started investing (Personal Communication external ML engineer/project manager, January 13, 2022). Besides the commercial evaluation, the marketeer evaluates the email expressions together with marketing communication advisors periodically, to see if they still agree on the messages that are being sent (Personal Communication Marketeer, January 14, 2021). Besides these evaluations on the commercial results, there is no periodic evaluation of the ML system itself or the sociotechnical ML system in place (Personal Communication Manager marketing intelligence, January 11, 2022). Furthermore, there is no continuous monitoring in place to detect potential changes in the ML system over time.

Conclusion of specification in the email marketing use case
In the email marketing use case, several elements of sociotechnical specification are identified. First, the use case started with defining the problem and analysing whether ML would be beneficial to be applied. Further, stakeholders with different expertise were involved in the specification of the use case. Additionally, requirements were defined by the different stakeholders. Furthermore, a validation period was initiated to validate the ML system in its application domain before operationalizing the system. Lastly, the ML system's commercial results are periodically evaluated with the direct stakeholders.

Despite these approaches, several potential shortcomings are identified in the sociotechnical specification of the email marketing use case. First, this bank is very new in the development and use of ML. Therefore, there are no guidelines or processes in place that could guide the project (Personal Communication marketing intelligence Manager, January 11, 2022). The specification in this use case was oriented towards the ML system and more importantly towards the commercial outcome. The total sociotechnical ML system also includes the human operator. However, the human operator was not consciously specified as part of the system, but was needed due to technical limitations. The interaction between the human operator and the ML system was no specific part of the specification. This could lead to vulnerabilities, as explained in Chapter 4. Furthermore, there was little emphasis on the specification of emergence in this use case. There are no monitoring or evaluation mechanisms specified for the ML system in use. Besides that, the interaction between the ML system and the human operator is not monitored or evaluated either.

5.7 Conclusion Chapter 5

The objective of Chapter 5 is to introduce the two use cases that are used to gather data about the application environment and to provide insight in the specification of the use cases. This leads to answering the third sub-question:

What does specification in the ML lifecycle look like in practice, based on two use cases in the financial sector?
Both use cases have elements that should be present in a sociotechnical specification of sociotechnical ML systems. However, other elements of sociotechnical specification are not present, resulting in shortcomings in the specification, possibly leading to vulnerabilities in the sociotechnical ML system. The current practice on specification is used as input to define the challenges from a sociotechnical ML systems perspective in Chapter 7. The next chapter, Chapter 6 dives deeper in the use cases to analyse how the identified dimensions of the theoretical framework are addressed in practice.

Chapter 6

Sociotechnical dimensions in ML practice

This chapter entails an analysis of the dimensions defined in Chapter 4 within the use cases and among civil society organisations and regulatory bodies. The objective of this analysis is to gain understanding in how the dimensions relate to the current practice of ML use case development. Insight into the challenges that are experienced in practice by different stakeholders, as well as potential gaps between what is understood from the knowledge base to be of importance and the awareness of these matters in practice. This is gained by investigating to what extent these dimensions are considered by stakeholders involved in the ML lifecycle of the two use cases, and whether vulnerabilities related to these dimensions actually emerged. Additionally, the perceptions of the civil society organisations and regulatory bodies broaden the view on the sociotechnical dimensions and provide a sense of relevance of each dimension. This leads to answering the fourth sub-question:

4. To what extent are sociotechnical dimensions addressed in practice, based on use case specific and general insights?

6.1 Deductive analysis of the interview data

To get insight in the issues in specification, development and use of ML systems that are encountered along the interviews, the interview data are analysed by means of an deductive analysis. This means that the coding frame has been developed at the beginning of the analysis process (Friese et al., 2018). This frame contained the dimensions defined in Chapter 4 as codes. During the analysis, the code frame has been enriched with additional codes to cover the content of the entire data set (Friese et al., 2018). This results in analysis on the level of the dimensions, which describes to what extent the dimensions are addressed in the use cases and what the perception on these dimensions is among the external interviewees.

After completing the analysis on dimension-level, another step has been taken to identify the challenges that emerge along all the dimensions. These are the challenges that serve as input from the application domain to the design cycle.

6.2 Misspecification

In both use cases, there exists awareness among the data science stakeholders that it is important to consider the larger sociotechnical system when developing an ML models. Nevertheless, examples of misspecification can be identified in both use cases. The misspecification dimension has been found relevant in practice among civil society organisations and regulatory bodies, which point out that technology is not the solution to every problem. Besides that, not specifying the larger sociotechnical ML system of which the ML system is part can lead to harmful consequences for organisations and individuals.

6.2.1 Misspecification in the use cases

In the financial crime detection use case, the manager of data science does realise that it is vital to take the role of the end-user into account during the development of the ML system, as the

value of the system would be zero if the end-user would not understand what comes out of the system. While the realisation is there, an example of misspecification can be identified in this use case. The description of the features to be used in the ML system were specified by the data science team, while the transaction monitoring analysts, being the end-users of the ML system output, have to use these descriptions for their analysis. This led to very technical descriptions, that were very difficult to understand by the transaction monitoring analysts. In the specification of the features, the end-user component of the sociotechnical ML system is thus not sufficiently considered, leading to a sociotechnical ML system that did not function properly. The transaction monitoring analysts had to make a translation document retrospectively to make the descriptions and thus the ML system output understandable for the analyst. This could have been prevented, by involving the end-user in the specification of the features and their descriptions.

Comparing this example of misspecification with the email marketing use case sheds light on another approach to deal with the specification of features, to prevent the misspecification described above. In this use case, the ML engineering/consulting firm intentionally chose not to use the most complex model and features. This choice was made to allow for the marketing intelligence analyst was able to understand the features and to propagate this to the rest of the organisation. The email marketing use case does have another example of misspecification. The ML engineering/-consulting firm initially automatically scheduled the model runs and transfers of the output files to the analysis system. However, the marketing intelligence analyst highlighted the importance of the checking role he has. Subsequently, it happened regularly that the transfer had failed, which led to a change in the process, where the marketing intelligence analyst now checks the model output before transferring to the analysis program. This example shows that a specification meant to lead to an efficient process is not necessarily desirable or even effective.

Moreover, in both use cases, the systems are specified for internal organisational objectives, not necessarily for fair outcomes for people affected by the systems. In the financial crime detection use case, the objective of one of the models is to reduce the false-positive rate of alerts being investigated. This reduction is seen as valuable for optimizing the alert investigation workflow. However, a false-positive alert means that a customer of the bank is investigated for financial crime, which is not mentioned as a problem. In the email marketing use case, the system is specified to increase the conversion rate towards investing products. This way, the system is optimized to select the most promising customers of the bank to start investing. This selection might lead to unfair outcomes, such as privileging those who are wealthy already to become even wealthier.

6.2.2 General insights on misspecification

Besides the identification of misspecification within the two use cases, the broader selection of interviews shed light on how misspecification can be present in ML use cases in general. First, several representatives of civil society organisations and regulators pointed out that technology is chosen as solution to problems, whereas it is not always the best solution. As the representative from Privacy First said: "What we see is a kind of love for technology, where goals are achieved with technological solutions, while solutions may need to be found in another area" (Personal Communication, January 22, 2022). Second, every model is a simplification of reality, and it is vital to understand that the world is more complex than the dots are ultimately present in an ML model, according to the representative of Bits of Freedom (Personal Communication, January 17, 2022). The fact that ML models are simplifications does not always penetrate the sociotechnical system, making ML be seen as a sort of holy grail, while it only captures something very specific, and is not a replacement for critical thinking in an organisation (Personal Communication representative Waag, January 19, 2022). Therefore, it should be debated whether developing an ML system is an appropriate solution to a problem, to prevent misspecification. Furthermore, there are risks of misspecification seen by several several representatives of civil society organisations and regulators in the usage of data. Data is often seen as something very factual and objective, while in reality it is a translation of what is seen in the world. Data is just used without really thinking and reflecting on the data and where it comes from, and what social problems have influenced that data (Personal Communication representative Bits of Freedom, January 17, 2022). If data is collected in one context, and later used for another context, this could lead to harmful consequences. An example here is that there were two data sets collected by two different organisations, that were later matched by a third party to find fraudsters, while one data set only the invoiced hours worked by a person, and the other data set contained the invoiced hours as well as the preparation time. This way two different specifications led to a misspecification in the use of the combined data,

causing people to be unjustly accused of fraud (Personal Communication representative Platform Bescherming Burgerrechten, January 18, 2022). Another consequence of misspecification could be that a model is developed and put into production, but the model is not used in practice. This is caused by a misspecification of the sociotechnical ML system of which the model becomes part. For example, the model was not needed, is not trusted by the end-user or adjustments in the way of working of the end-user are not secured (Personal Communication Advisor Data Science, December 22, 2021).

6.3 Machine error

Machine error can be roughly divided into two categories; machine incorrectness and machine bias, as explained in Section 4.4.2. The impact of decisions made based on the models in the two use cases is fairly different. This seems influential to how is dealt with the potential machine error in the models. In the financial crime detection use case, the performance of the models' internal workings are always evaluated before pushing the output to the analysts, whereas in the email marketing use case the model is trusted to be working as expected. Machine error is recognised by the civil society organisations and regulatory bodies as a relevant dimension.

6.3.1 Machine incorrectness in the use cases

In the financial crime detection use case, the models are used in a quite sensitive context, which makes it important to prevent incorrectness in models. The impact of incorrectness could be that customers of the bank are unjustly investigated by the transaction monitoring analysts, or that customers that should be detected by the model are not detected, which could lead to money laundering or terrorist financing being undetected. Because the latter is considered to be important to prevent, there is an acceptance of a less accurate model, which leads to more false positives, in order to be able to find the true positives. Further, there are multiple mechanism in place to prevent incorrectness in this use case. First, data scientists work with a four-eyes principle, by which every piece of code is checked by another data scientist during development. Second, there is a dedicated independent model validation team that validates a model including all code before it is put to production. Lastly, there is monthly performance monitoring in place for the models, that run monthly, to check whether the model performance is comparable to its performance during training and whether the features' distributions has changed to detect potential machine incorrectness. After completion of the performance monitoring, the model output is directed to the transaction monitoring analysts. Although the bank has these mechanisms in place, the interviewed transaction monitoring analyst pointed out that in a third model, of which the first version is currently live, the testing of the model was not performed in the case management system the analysts are using. Once the first real output of the model in the case management system appeared, there was a lot of incorrectness; for example, generated alerts that did not contain transactions and features that were not visible in a customers' account. As a result, the analysts are dealing with model output that has a lot of incorrectness, and a new version is still not live at the time the interview was conducted (Personal Communication, January 14, 2022).

On the contrary, every interviewee in the email marketing use case pointed out that the impact of machine error is relatively low. The model is not part of a mission-critical activity, and the worst case impact to customers is that customers are accidentally repeatedly being emailed by the bank, or are being emailed while they had opted-out for certain emails. Although this relatively low impact, the marketing intelligence manager mentions the importance of being transparent. For example, if customers that had opted-out for emails would have been emailed (Personal Communication, January 11, 2022). In the first live version of this model, there appeared to be overfitting on a certain feature. This was only detected once the model was already live, because the test set did not contain the needed exceptional cases. There was only 1.5 years of data so most of the data had to be used for training, but if more data had been available for testing this overfitting could have been detected earlier, before going live (Personal Communication external ML engineer/project manager, January 13, 2022). To prevent emails being sent based on incorrect model output, there is a human in the loop, the marketing intelligence analyst, who does manual checks on the model output to check for example is customers had opted-out. Although he checks whether proposed customers on some standard characteristics, he trusts the model to be working as expected (Personal Communication marketing intelligence analyst, January, 21, 2022).

6.3.2 Machine biases in the use cases

To prevent biases in the models, in both use cases it was chosen not to use certain data. In the email marketing case, the external ML engineer/project manager pointed out that the bank thinks it is very important to use data wisely, well within the lines of the GDPR (General Data Protection Regulation). Therefore, the choice was made not use gender data and postal codes (because those may be a proxy for ethnicity). The choices of what could not be included were made based on intuition. In the financial crime detection use case, the data science department wanted to detect potential bias in the ML models caused by proxies for gender, age, ethnicity and origin. However, the privacy office prohibited to do this for ethnicity and origin, because those are sensitive personal data. This presents a notable paradox: To be able to detect certain bias on a particular factor, you have to use data on this particular factor, for example ethnicity. However, by the GDPR, an ML model should adhere to 'privacy by design', which means data on ethnicity and ethnicity cannot be processed, and therefore not detected in this way.

Reflecting on the handling of potential bias in the ML systems in the use cases, not using certain features and detecting bias for certain factors by proxies do not directly lead to fair outcomes. This perspective on biases does not account for the diverse fairness requirements and needs stakeholders involved in the sociotechnical ML system might have (Balayn & Gürses, 2021).

6.3.3 General insights on machine error

Machine error was recognized as a relevant dimension by the representatives of the civil society organisations and regulators. To illustrate, biases are seen as one of the largest problems in ML systems, that predict behaviour or crime in particular, by Amnesty International. A large risk is the use of indicators that are actually proxies for protected grounds, as was discussed within the use cases as well (Personal Communication, January 25 representative Amnesty International, 2022). The representative of the AP also points out to guard for unwanted indirect inferences to protected grounds, although it is increasingly difficult to completely prevent this from happening because of the ever-growing amount of data (Personal Communication, January 24, 2022). There is a lot of attention on the impact of machine error on people in the proposed EU AI act (Personal Communication representative DNB, January 18, 2022). Society does not accept machine error, which makes the incentive for organisations to be open about it not that great. This non-acceptance reflects in a statement made by the representative of Amnesty International, who stated that ML models with large error margins which decisions can have large consequences for people, should not be used in practice (Personal Communication, January 25, 2022).

6.4 Interpretation

Interpretation of an ML system by human decision-makers is recognised by many civil society organisations as a potential source of risks for the quality of the ultimate decisions. At the same time, human intervention is mandatory if a decision affects a person to a significant degree in the GDPR. In the use cases, stakeholders seem to be less aware of the vulnerabilities of human intervention. In the marketing use case, human intervention is seen as a means to reduce risks by providing an extra checking and controlling mechanism, instead of a step that can impose additional vulnerabilities. There is an increased risk of human noise in the financial crime detection use case, as the more complex system output increases the potential for deviation in interpretations among analysts. Further, the potential for human bias is not adequately considered in both ruse cases and the human interpretation step is not monitored or evaluated.

6.4.1 Interpretation in the use cases

In the financial crime detection use case, the analysts that use the model output in their work encountered a challenge in the beginning to use the model output. The number of factors that are taken into account using the ML models instead of the previous rule-based system was largely expanded, which made it difficult to understand how to interpret the output, where to look at and what to think about at all (Personal Communication transaction monitoring analyst, January 14, 2022). Meanwhile, the transaction monitoring analyst thinks the model output is well interpretable, in combination with the translation document on the model features. The analysts are supported in the model interpretation by means of explainability methods, consisting of highlighting the three most important features that contributed to the model output, as well as the

most important transactions (Personal Communication Lead Data Scientist, January 12, 2022). Further, analysts that start working with the ML models' output follow a training and receive a working instruction document. Nevertheless, the transaction monitoring analyst pointed out that the introduction of the ML models increases the risk of deviations in interpretation among analysts compared to the previous rule-based system, because the output is a lot more complex (Personal Communication, January 14, 2022). The advisor data science within the bank does see that bias could emerge in the decision-making process besides the model, but that it is difficult to quantify this (Personal Communication, December 22, 2021). The manager data science points out that there are a lot of checks on the model itself, but potential human bias is not very well considered, which could be improved. At the same time, the human intervention is an important mitigation measure for automated decision-making, and completely automating the decision-making is not desired for these important decisions (Personal Communication manager data science, January 12, 2022).

In the marketing case, a human-in-the-loop, the marketing intelligence analyst, has a controlling and checking responsibility for the model output, after which the proposed customers by the model are being sent an email. The model output only presents the customer IDs that are proposed, which makes it difficult to understand why the model chooses certain customers (Personal Communication marketing intelligence analyst, January 21, 2022). The marketing intelligence analyst feels that the checking function he has is very important, however, he does not have enough tools and guidance to perform this function adequately. He would like to have better reports on the customers that are selected, to get a picture of the customers (Personal Communication marketing intelligence analyst, January 21, 2022). The ML engineer/project manager does not see a great risk for human bias to emerge, as the marketing intelligence analysts mainly has a checking role (Personal Communication, January 13, 2022). Additionally, the marketeer thinks the role of the marketing intelligence analyst rather decreases the risk on potential mistakes, than imposing new potential mistakes (Personal Communication, January 14, 2022).

6.4.2 General insights on interpretation

Both representatives of Bits of Freedom and Amnesty International mention that a human-in-the-loop is seen by many organisations as a means to take away concerns about ML, while it is not the solution to the problem and much more is needed as prerequisites that a human-in-the-loop (Personal Communication, January 17 and 25, 2022). Automation bias and limited time for the job can make humans overly rely on the model output, and a model can be used by humans as confirmation to their own bias (Personal Communication representatives Platform Bescherming Burgerrechten and Amnesty International, January 18 and 25, 2022). At the same time, human intervention in automated decision-making that can affect a person to a significant degree is mandatory by the GDPR (Personal Communication AP, January 24, 2022). The ML engineer within bank A points out that the chain of activities in a decision-making process around an ML model is more important for the quality of the ultimate decisions than the model itself (Personal Communication, January 24, 2022). It really depends on the mindset of the human decision-makers that are in the chain between the model output and the final decision. If the decision-making process is badly designed or not thought through, a very good model can lead to terrible outcomes (Personal Communication ML engineer bank A, January 24, 2022).

6.5 Behaviour

Behaviour as a dimension is not largely represented in the interview data about the use cases, which could be caused by unawareness of the dimension or behaviour being of less relevance in these use cases. The civil society organisations and regulatory bodies did recognise behaviour. The representative of DNB pointed out that it is very relevant, yet the point of concern does not get much attention in the field, which could be an explanation of the dimension not being extensively elaborated on in the use cases.

6.5.1 Behaviour in the use cases

Considerations on behaviour have not been mentioned in the email marketing use case. In the financial crime detection use case, there was one comment on behaviour. The ML models required change in the way of working of the analysts, to which they reacted that they could not work or

did not want with the model output (Personal Communication Manager Data Science, January 12, 2022).

6.5.2 General insights on behaviour

The general insights on behaviour consist of insights on the change of behaviour among people about whom a ML model takes a decision, as well as change of behaviour among human decision-makers that use the model output to take a final decision. The representative of DNB argued that a trade-off exists between transparency and black-box models. For example, if a criminal gets insights in how an ML model that is used to detect money laundering works, become easier to circumvent being detected as a money launderer. This point of concern get little attention and does not play a big role in the discussion around explainability (Personal Communication, January 18, 2022). The representative of Amnesty International sheds a different light on this trade-off, arguing that it is not required to make the whole code public, but people should be informed when personal characteristics as nationality, age, postal code or salary are used. The potential for circumventing the system is not at stake, as people cannot change these characteristics, but should be able to know based on what a decision is made (Personal Communication, January 25, 2022). behavioural change among human decision-makers that use model output is also mentioned by interviewees. The introduction of an ML model decreases the level of ownership a human decision-maker has, compared to a situation without an ML model in place, which can influence the outcome (Personal Communication representative of Waag, January 19, 2022). Lastly, if a human decision-maker has to adhere to a predefined target in terms of validating and rejecting model output, this has influence on the behaviour and the final decisions as such (Personal Communication AP, January 17, 2022).

6.6 Adaptation

Adaptation as a dimension can be recognised in the use cases, although it has not been widely discussed in the interviews. In the financial crime detection, the adjustments in the way of working were very challenging, which was not an issue in the email marketing use case. In that use case, the marketing intelligence analyst has actively been given the possibility to adjust configurations towards the environment's needs. Only one representative of a civil society organisation had encountered examples of the adaptation dimension in practice. Whereas it seems that adaptation is not on top of mind among the external stakeholders, ML systems were developed and not being used in practice due to the adaptation of the designated end-users of ML systems in Bank A.

6.6.1 Adaptation in the use cases

In the financial crime detection use case, the transaction monitoring analyst pointed out that it had been a big challenge to start working with the ML models, as it required a whole new way of working (Personal Communication, January 14, 2022). For the transaction monitoring analysts to understand what the models' output meant in the context of actual increased risk of money laundering or terrorism financing was a struggle in the beginning of using the models in practice To address the difficulties of using the models' output in practice, there have been intensive feedback sessions between the data science team and the transaction monitoring analysts (Personal Communication transaction monitoring analyst, January 14, 2022).

In the email marketing use case, the marketing intelligence analyst has been given the possibility to adjust some configurations of the model. For example, he can adjust the threshold of the model output, by which a change leads to more or less customers to be emailed based on the model output (Personal Communication external ML engineer/project manager, January 13, 2022). While these possibilities give the marketing intelligence analyst the possibility to adjust the system to the needs of the environment, he has been given instructions not to change configurations too often, because it makes the ML system hard to evaluate (Personal Communication external ML engineer/project manager, January 13, 2022).

6.6.2 General insights on adaptation

As the introduction of ML models often require a change in the way of working among the human decision-makers that are part of the decision-making process, this should be accommodated for in the development and integration of the ML system in its context. In bank A, there have been models

that were developed and put into production, but were not used very much in practice, because this accommodation was lacking or there was no trust in the model (Personal Communication advisor data science, December 22, 2022). Another phenomenon within the adaptation dimension is function creep: It happens a lot that an ML system that is designed and developed for one purpose is over time used for other purposes as well (Personal Communication representative Amnesty International, January 25, 2022). Moreover, it can happen that the initial design of the system does not show risks on human right violation, but the way in which the system is used in practice does, for example leading to groups of people being treated differently (Personal Communication, January 25, 2022).

6.7 Dynamic change

Dynamic change is not widely recognised by the civil society organisations. Most representatives did not encounter an example of dynamic change in practice, but can image the relevance of it. On the other hand, stakeholders involved in the use cases do recognise this dimension, as external factors such as regulatory changes or internal changes in data could have large impact on the working of the models over time, while alignment of changes within the bank's different departments involved is a challenge.

6.7.1 Dynamic change in the use cases

In both use cases, the department in which the models are developed and ran in production are not the owners of the underlying data sources. At the same time, changes in underlying data can have direct impact on the models. In the financial crime detection use case, the Manager Data Science pointed out that if a new version of an ML model is developed, this requires all governance checks, while the IT department that is the data owner changes data sources, these governance checks are not in place (Personal Communication, January 12, 2022). Therefore, the data science team needs to continuously align with the IT department about data changes. Additionally, they monitor potential changes in data or output that can be indicative of problems in the models, monitoring metrics such as feature distributions, alert volume, and percentage of false positives are used (Personal Communication Lead Data Scientist, January 12, 2022).

In the email marketing use case, changes in underlying data are seen as a relevant dimension by all stakeholders involved. At the same time, changes are not always adequately communicated to the marketing intelligence department, while significant changes could result in no output from the model (Personal Communication Marketing Intelligence Analyst, January 21, 2022). Other developments within the bank, such as updated privacy guidelines or new products, could require adjustments in the model as well. Currently, the department is depending on the external ML engineering/consulting firm to make adjustments in the model (Personal Communication Marketing Intelligence Analyst, January 21, 2022). Lastly, external factors could really impact the model performance. For example, the stock exchange dip in March 2020 due to the start of the Covid-19 pandemic in The Netherlands, was a scenario the model has not been trained on, which leads to unreliable output that should not been used (Personal Communication external ML engineer/project manager, January 13, 2022).

6.7.2 General insights on dynamic change

Representatives of Waag, Bits of Freedom, Amnesty International and Platform Bescherming Burgerrechten have not encountered dynamic change as a dimension in practice, but can image that it is a source for vulnerabilities. On the other hand, the representative of DNB does recognise the problems for supervision that dynamic change could cause. The proposal of the new AI act does require a conformity assessment for new high risk applications of ML, but does not take the dynamic change dimension into account (Personal Communication representative DNB, January 18, 2022). Especially in the case of self-learning ML models, an investigation on one day could lead to the outcome that the model is compliant, but the next week it could not be compliant any more (Personal Communication representative DNB, January 18, 2022). This is a realistic problem that challenging for supervising bodies.

6.8 Downstream impact

Among representatives of civil society organisations and regulatory bodies, downstream impact is seen as an important dimension, in which especially data quality and data selection can have severe impact on the outcomes of ML models. At the same time, the subject receives little attention in practice compared to the attention to the development of the ML models themselves, which has been noticed in the interviews with stakeholders involved in the use cases, in which the concerns of downstream impact seemed to not play a large role.

6.8.1 Downstream impact in use cases

Within bank A, where the financial crime detection use case was developed, there are measures to ensure good data quality and compliance with the GDPR, as the bank mostly uses gold standard data sources for ML models, which are the most accurate and reliable of its kind (Personal Communication Privacy Officer, January 17, 2022). At the same time, the data sources are maintained and continuously improved or changed by the IT department, which can have downstream impact on the models (Personal Communication Manager Data Science, January 12, 2022). A large challenge is to keep grip on where ML is used within the organisation (Personal Communication Privacy Officer, January 17, 2022). To keep a grip on where model output is used and thus limit downstream impact, employees or department need either a data sharing agreement or authorization from the data owner, who is the data scientist who developed the model, to use the model output for different purposes (Personal Communication ML engineer, January 11, 2022). It is clear that measures have been taken in bank A to prevent vulnerabilities due to downstream impact. However, the changes in data sources are important to monitor, as well as the requirements for departments to be able to use model output in secondary decision-making processes.

In the email marketing use case, the occurrence of vulnerabilities within the downstream impact dimensions seems limited, as the data quality was good in general, and the output of the ML model is not used in secondary decision-making processes. However, the marketing intelligence analyst did see the possibility for this to become the case in the future (Personal Communication, January 21, 2022). Downstream impact is thus a dimension to keep in mind in bank B.

6.8.2 General insights on downstream impact

Downstream impact is seen as an important source of vulnerabilities by civil society organisations and regulatory bodies. Both downstream impact by data issues as downstream impact by interconnected ML models have been mentioned. The representative of DNB called data management, governance and quality at least as important as the development of models, whereas it is a relatively small part of the discussion (Personal Communication, January 18, 2022). Representatives of Amnesty International and Bits of Freedom also highlighted that the data used can have severe impact on model output, which receives too little attention (Personal Communication, January 17 and 25, 2022). An ML model can in turn interact with other ML models, which can lead to losing control on wrong model output or if one model fails, other models fail too, which can have large impact on the larger system (Personal Communication representatives DNB and Waag, January 18 and 19, 2022).

6.9 Accountability

Lack of accountability can lead to large issues as outcomes based on ML models can have big impact on people, is a shared point of view from the civil society organisations and regulatory bodies. Within the use cases, there can be vulnerabilities related to accountability identified. In the financial crime detection use case, remaining knowledge on the models within the bank is challenging, which is key to be able to provide accountability. In the email marketing use case, reproducibility of the customers selected and dismissed is not easy to achieve, and the responsibilities are not officially defined among the involved stakeholders.

6.9.1 Accountability in use cases

In the Financial Crime Detection use case, accountability of the model, model output and outcome are divided among different stakeholders. The data science team is model owner and thus responsible for the model. The ML team carries responsibility for correct implementation of the model,

and the transaction monitoring analysts are responsible for the decision whether a customer should be reviewed or not. Ultimately, the leader of the financial crime detection department carries end-responsibility for everything that happens in the department, and thus for the model, model output and final outcomes (Personal Communication Lead Data Scientist, January 12, 2022). To be able to provide accountability over the ML models used for certain output, reproducibility is in place for the model version, model output, features and optionally parameters, metrics and other metadata (Personal Communication ML engineer, January 24, 2022). Being able to explain an outcome to a customer is a challenge within bank A, because there is a continuous flow of employees leaving and joining the bank (Personal Communication Privacy Officer, January 17, 2022). The risk exists that at a certain point, nobody knows how an ML model works any more. To prevent this, transparency on model development is key (Personal Communication Privacy Officer, January 17, 2022). As the model output is used by the transaction monitoring analysts in the use case, they have to work uniformly to make the final outcome reproducible as well (Personal Communication Transaction Monitoring Analyst, January 14, 2022). It is vital to be able to explain how a model works and how the outcome is achieved to the Autoriteit Persoonsgegevens if asked(Personal Communication Privacy Officer, January 17, 2022). As the model output is part of a larger decision-making process, making model output explainable is not enough to being able to understand the final outcome (Personal Communication ML engineer, January 24, 2022).

Being able to explain to a customer why he or she has been selected is seen as a great challenge in the use of ML among stakeholders in the email marketing use case (Personal Communication Manager marketing intelligence and Marketing Intelligence Analyst, January 11 and 21, 2022). If a customer asked this, the bank cannot fully explain why he or she receives the email (Personal Communication Marketing Intelligence Analyst, January 21, 2022). At the same time, customers could always unsubscribe from receiving certain types of emails (Personal Communication Marketing Intelligence Analyst, January 21, 2022). Also, the model uses only seven features, so the stakeholders within the bank have insight into the data that have led to a model output, which makes the model explainable to a certain degree. The model has been developed by an external ML engineering/consulting firm, but the responsibility for using the model and the model output is carried by the bank (Personal Communication external ML engineer/project manager, January 13, 2022). Within the bank, the responsibilities of the model and model output are not officially defined (Personal Communication marketeer, January 14, 2022). Reproducibility of model output is covered by the external ML engineering/consulting firm. However, reproducibility of the final outcome, thus customers that are selected and the customers that are dismissed are not easy to reproduce, because the final selection of the customers is overwritten every week. If needed, the marketing intelligence analyst could match the model output data with the customers that have been emailed to trace back the final selection (Personal Communication Marketing Intelligence Analyst, January 21, 2022).

6.9.2 General insights on accountability

The lack of accountability of ML systems is seen as a big issue among civil society organisation and regulatory bodies. Issues related to accountability can be divided into internal accountability issues and external accountability issues. Internal accountability issues are issues within organisations. In the financial sector, the board level is ultimately accountable for the use of ML models within the organisation, whereas they often do not fully understand what the use of ML entails due to a lack of knowledge and skills (Personal Communication representative DNB, January 18, 2022). There are very few experts within and outside of banks that understand how ML models really work. This could lead to board members not being aware of the risks and the impact of ML models on the organisation and the larger financial system (Personal Communication representative Waag, January 19, 2022). External accountability issues entail that people that are affected by a decision made using an ML system that decisions can not be explained and people do not have the channels to defend themselves to seek justice. These problems are partly caused by the secretive way of working among organisation. While the GDPR requires organisations to be transparent about the ML models that are used, the AP sees lack of transparency and proactive communication among organisation (Personal Communication representative AP, January 24, 2022). When people are not informed what is going on, it is impossible for them to detect errors in the outcome (Personal Communication representative Platform Bescherming Burgerrechten, January 18, 2022). As this outcome can have large impact on individuals or groups of people, the outcomes should be explainable, and the user should be held accountable, which is often not the case in practice (Personal Communication representative Bits of Freedom, January 17, 2022).

6.10 Conclusion Chapter 6

This section aims to answer the fourth sub-question:

4. To what extent are sociotechnical dimensions addressed in practice, based on use case specific and general insights?

Chapter 6 presents a deductive analysis of how the eight dimensions defined in 4 are addressed in the financial crime detection use case and the email marketing use case. Additionally, the perspectives of the civil society organisations and regulatory bodies provide additional insights in to what extent the dimensions are addressed in general.

Both use cases show efforts to address certain dimensions. For example, in both use cases, a four eyes principle is being used to prevent machine error in the ML models. Besides that, performance monitoring is used in the financial crime detection use case to detect dynamic change on the ML system level. Furthermore, the specification of the email marketing use case takes the sociotechnical context into account by using understandable features for the organisation.

Despite the efforts made, the potential for vulnerabilities to emerge within the sociotechnical dimensions can be identified in both use cases. For example, in the behaviour and adaptation dimensions, there has been identified a lack of awareness in the use cases. Dynamic change does not have priority in the email marketing use case, as no measures are taken to detect or prevent dynamic change. Lastly, the application context of the ML systems was insufficiently considered in the financial crime detection use case, as features were described in an incomprehensible way for the transaction monitoring analysts that have to use these descriptions in their investigations.

Moreover, there is a lack of awareness of the larger implications the ML systems might have on customers of the banks. In both use cases, the objective is to optimize internal goals of the banks, rather than optimizing the outcomes for customers. Fair outcomes are not a large point of consideration. For example, the potential of discriminatory outcomes was not considered in the email marketing use case, as not using certain data was seen as enough prevention of discriminatory outcomes. Also, the selection of customers is based on the potential conversion rate, which might lead to the selection of customers that are already privileged to become even richer, ignoring other customers that have a smaller predicted conversion rate, but could benefit from starting investing. In the financial crime detection use case, reducing false positive alerts on potential criminal customers is seen as a means to optimize the investigation workflow. The fact that false positive alerts mean that customers of the bank are unjustly investigated for financial crime is not mentioned as problematic in itself. Moreover, the possibility that those false positive alerts might be biased, resulting in certain groups of customers being investigated more than others, plays a limited role.

The insights gathered in this chapter will be used as input to define the main challenges in ML practice from a sociotechnical ML systems perspective in Chapter 7.

Chapter 7

Challenges in ML practice

This chapter presents the most important challenges identified along the dimensions within the use cases and the insights provided by the representatives of the civil society organisations and regulators. This chapter aims to answer the fifth sub-question:

What are the main challenges identified in ML practice, seen from a sociotechnical ML system perspective?

7.1 Inductive analysis of the interview data

Where Chapter 6 presents the deductive analysis of the dimensions in practice, this chapter presents an inductive analysis of the interview data. An inductive analysis is data driven, and the researcher is not trying to fit the data into a pre-existing coding frame (Friese et al., 2018). This analysis allows the researcher to analyse along the dimensions and additional interview data to identify the main challenges in ML practice from a sociotechnical ML systems view. Table 7.1 presents an overview of the main challenges identified and how they are associated to the dimensions of the theoretical framework. As can be seen, all dimensions are presented, which means that insufficiently addressing the challenges can lead to vulnerabilities in the associated dimensions.

Table 7.1: Overview of the challenges and associated dimensions

Challenge	**Associated dimensions**
Challenge 1: Defining the system boundaries for the ML lifecycle	Misspecification, Interpretation, Behaviour
Challenge 2: Dealing with emergent dimensions in the sociotechnical ML system	Behaviour, Adaptation, Dynamic change, Downstream impact
Challenge 3: Introducing a human decision-maker in an ML decision-making process	Misspecification, Interpretation, Behaviour, Adaptation
Challenge 4: Recognising the importance of data in the ML lifecycle	Misspecification, Machine error, Dynamic change, Downstream impact
Challenge 5: Developing knowledge and communication within sociotechnical ML systems	Misspecification, Machine error, Dynamic change, Accountability
Challenge 6: Providing transparency of ML systems and outcomes	Interpretation, Behaviour, Downstream impact, Accountability
Challenge 7: Interpreting regulations applicable to ML systems	Interpretation, Dynamic change, Downstream impact, Accountability

7.2 Challenge 1: Defining the system boundaries for the ML lifecycle

In specification of the ML lifecycle, the system boundaries need to be defined. This defines which components and interactions are taken into account in the ML lifecycle, and which stakeholders are involved as such. When the system boundaries are defined too narrow at the start, this has consequences for the efficiency and effectivity of the system development and is a source for vulnerabilities.

As the objective in the ML lifecycle is to develop an ML model and put it to production, a technical view on the ML system is needed and used in both use cases to develop the model, perform model testing and model validation. To prevent and detect potential machine error, a four eyes principle is used in both use cases in the model development. Further, in bank A, an independent model validation team takes the whole model development, choices made and final model under the loop to validate the model. After validation, the model is put to production and is used to make predictions.

However, an ML model does not operate in isolation, but becomes part of a larger sociotechnical ML system. Not including this within the system boundaries may lead to issues later in the ML lifecycle. Firstly, the end-users of the ML system, the transaction monitoring analysts, were not involved in the financial crime detection use case from the start. They were not involved until the model had been developed and was ready to be tested and become part of the work of the end-user. For the end-user, the introduction of the ML model required a great shift in their way of working. Moreover, they were not able to interpret the model output, as the feature descriptions they had to use were defined in technical language. Therefore, a translation of the feature descriptions had to be specified ad-hoc and working instructions needed to be created by a delegation of the transaction monitoring analysts. Secondly, as described in the previous paragraph, bank A has many governance processes in place to safeguard the introduction and use of ML systems. However, these are focused on the ML system, and not on the larger sociotechnical ML system, in which issues could be present and remain undetected. Thirdly, a model within the transaction monitoring use case had not been tested in the actual application domain before it was put to production. This led to many errors in the alerts the transaction monitoring had to investigate.

Further, the larger implications of the introduction of an ML system should be considered within the system boundaries. ML systems can have discriminatory or exploiting impact on (groups) of people. It is vital to consider these larger implications to establish fair outcomes of sociotechnical ML systems for people.

Ultimately, if the larger sociotechnical ML system is not adequately considered in the ML lifecycle, this could result in ML systems being developed, but ultimately not being used in the organisation due to a lack of accommodation in decision-making processes, lack of trust, or no alignment with the business goals. As such, organisations spend much time, money, and effort in the development of ML systems that ultimately do not bring value to the organisation. Consequently, the belief in the value of ML may vanish, which blocks future ML initiatives in the organisation.

As illustrated, defining the boundaries too narrow is a source of issues. On the other hand, it is not feasible to involve everyone and specify every detail within the sociotechnical ML system from the beginning. As such, defining adequate system boundaries is a challenge.

7.3 Challenge 2: Dealing with emergent dimensions in the sociotechnical ML system

As concluded in challenge 1, not every detail within the sociotechnical ML system can be specified in the beginning. This is especially the case because vulnerabilities may emerge over time, emerge due to interactions, and emerge due to dynamics within the sociotechnical ML system.

To deal with emergence over time, several efforts are made in the use cases. Firstly, to keep a grip on the ML model's performance over time, monitoring metrics have been defined by the data scientists in the financial crime detection use case. For example, accuracy, prediction volume and

feature distributions are used to monitor whether the model performance remains similar as during development. This way, dynamic change can be detected and anticipated on. Secondly, feedback from the end-users is collected in both use cases to improve the ML models over time. Thirdly, a periodic evaluation is in place with the most involved stakeholders to evaluate the commercial results accomplished by the ML system in the email marketing use case. Fourthly, within bank A, models have a period of validity, after which extensive evaluation or deployment of a new model version is required. Lastly, the investigation reports the transaction monitoring analysts write using the ML system output is regularly controlled within the operations department.

In the email marketing use case, there is no performance monitoring in place. The model is now part of 'business as usual' and it trusted to perform as expected. Consequently, if vulnerabilities emerge in the ML model and its output over time, those are potentially not detected. Furthermore, the human operator of the model has the possibility to adjust configurations on the model, which could accommodate adaptation. At the same time, the human operator is discouraged to do so, because it hampers the evaluation of commercial results.

Furthermore, there are blind spots identified in the use cases on emergent dimensions. First, the behaviour dimension is hardly addressed in both use cases. This could for example lead to human decision-makers that overly rely on ML system outputs due to time pressure. Second, the adaptation dimension is insufficiently considered, as the way of working was not part of the specification in both use cases. Third, the feedback gathered from end-users is used to improve the ML systems. However, this does not take into account that human bias or noise can be inferred by this feedback.

As illustrated, vulnerabilities may emerge after deployment, when an ML system is used in its sociotechnical context. Awareness of the emergent dimensions is needed, and how to address them is challenging.

7.4 Challenge 3: Introducing a human decision-maker in an ML decision-making process

A human intervention in a decision-making process is mandatory if automated decisions affects a person to a significant degree, as described by the GDPR. However, a human-in-the-loop brings about new potential vulnerabilities that can lead to harmful outcomes of an ML system, such as automation bias, confirmation bias, and limited time for the job. At the same time, these vulnerabilities are not adequately addressed in practice. In the email marketing use case, the human operator is seen as reduction of potential errors, instead of causing new errors and biases. However, the human operator is not provided with enough guidance to be able to interpret the model output and detect potential errors.

In the financial crime detection use case, human decision-makers should be able to deliberately use the model output for their final decision. To do so, they have to understand the model output to that extent. Initially, human decision-makers experienced difficulties to understand the model output and use it in their work. Eventually, this is addressed using explainability techniques, to guide the human decision-makers in using the model output. This helps them, but the introduction of the ML system in their work has increased the risk of deviations in interpretation, as the complexity of information has grown.

Furthermore, the introduction of an ML system can impose vulnerabilities in the behaviour and adaptation dimensions. On both dimensions, awareness in lacking in the use cases, as these dimensions did not come forward in the interviews.
Although the introduction of a human decision-maker is often mandatory by the GDPR in an ML decision-making process, it must not be seen as a means that takes away vulnerabilities. Rather, it may introduce new ones. As the current focus in the specification of use cases is mainly on safeguarding the ML system itself, the specification and design of the decision-making process it becomes part of has less priority. As the ML engineer within bank A summarized: "you can have a very good model, but if the decision-making process around it is badly designed, a very good model can lead to terrible outcomes."

7.5 Challenge 4: Recognising the importance of data in the ML lifecycle

As data is at the core of ML systems, it has large influence on the final output of the model and the potential for vulnerabilities to emerge. Biases in data, bad data quality, and changes in data are thus important vulnerabilities to consider.

First, data is not objective. It is a translation of what is seen in the world, what social problems are there such as discrimination, which makes it important to cautiously use data. The GDPR offers guidance to organisations on personal data usage, and may direct organisations not to use certain personal data of persons. Although these data are not used, other data in the data set can still indirectly represent the initial data. For example, postal codes can indirectly represent certain ethnic backgrounds. This possibility of so-called proxies are recognised in both use cases. In the email marketing case, the choice have been made to not include postal code to prevent this from happening. The choice on what data to include and what not are made based on the intuition of the ML system developers, which raises the question if this intuition led to just choices. In the financial crime detection use case, the possibility for bias in data by proxies has also been recognised. The data science team wanted to get insight into potential bias on age, gender, ethnicity and origin. However, due to the GDPR, this was prohibited for ethnicity and origin. This presents a notable paradox: to be able to detect and eliminate bias by proxies in the ML systems, the actual data on ethnicity and origin has to be processed, which is prohibited. As such, organisations cannot detect bias in this way.

As the representative of DNB pointed out: "data management, governance and quality is at least as important as the development of models, whereas it is a relatively small part of the discussion". This statement is recognised in the use cases by the researcher. In both banks, data management lies in a different department than where the ML systems are developed. In bank A, the data science department has to follow very strict governance processes to be able to launch a new model or model version. However, the data sources are not covered by these strict governance processes. As such, changes in data can be made, which directly influences the ML models and output. Keeping a grip on this, requires continuous alignment between the departments. The latter is also recognised in the email marketing use case. This organisational complexity can be a source of vulnerabilities when changes are not communicated within the organisations.

7.6 Challenge 5: Developing knowledge and communication within sociotechnical ML systems

As different types of knowledge exists and are needed within sociotechnical ML system, they have to be developed, shared and maintained. On top of that, stakeholders with different types of knowledge should communicate to develop and operate an adequate sociotechnical ML system

Firstly, there are very few experts on ML within banks and in external organisations. This has the following consequences that could result in vulnerabilities. First, Bank A does have ML experts in-house, but ML experts often leave the bank and new ML experts are hired. This makes maintaining knowledge on models that run in production a challenge. To address this, extensive documentation on the models is made and updated. Second, bank B does not have ML experts in house. In the email marketing use case, bank B hired an external ML engineering/consulting firm to develop the ML system. Consequently, the bank is dependent on the external party if changes in the ML system have to be made due to emergent dynamics.

Secondly, the expertise of ML experts and the expertise of end-users on the way of working lies far apart. In both use cases, the people that ultimately have to work with the ML system output are not ML experts. The other way around, ML experts are no operational experts. As such, to develop an ML system that can be integrated in the way of working of operational experts, communication between the different stakeholders is needed for alignment. However, this communication is often lacking in organisations. To illustrate with the financial crime detection use case, the operational experts were only involved when the ML system development was finished. Also, the model validation team only validated the models, and did not communicate with operational experts to validate whether the models could actually be integrated in the way of working.

Lastly, as using ML is in its infancy in many organisations, governance, internal guidelines, and policies are often simply not existing. Knowledge and awareness on what it entails to develop an ML system and what potential vulnerabilities it can bring about are lacking within organisations. This can be recognised in the bank B. In the email marketing use case, the data used had to be approved by a compliance officer, and the department has to undergo a general risk assessment twice a year. However, no specific bank-wide guidance for ML development and use exists, which makes it possible for vulnerabilities to emerge.

7.7 Challenge 6: Providing transparency of ML systems and outcomes

Transparency of ML systems is challenging on multiple levels. First, ML models are known for their opacity, also referred to as a 'black-boxes'. Therefore, it is challenging to understand why an ML model takes a decision. To address this opacity, explainability techniques are used in the financial crime detection use case. This provides the end-users guidance in understanding the model output. In the email marketing use case, the complexity of the model itself is kept low and just a few features are used, to provide an interpretable model to a certain extent. This makes it always possible to analyse the features used for a prediction to get insight. However, it is still not possible to completely understand the model output for the stakeholders involved. The opacity of the ML systems complicates the ability to detect vulnerabilities in the systems.

Second, transparency and being able to explain decisions to customers is seen as the largest challenge in the development and use of ML by the manager marketing intelligence in bank B. If organisations are not transparent about the use of ML and how decisions are made, it is difficult for customers to defend themselves to decisions. Generally, organisations are lacking transparency on the use of ML systems for decision-making appeared from the interviews with civil society organisations and regulators. As a result, people are unable to detect and address potential fundamental rights violations or unjust decisions in sociotechnical ML systems.

Thirdly, organisations that use ML are often non-transparent about it towards civil society organisations, regulators, and other organisations. As a result, civil society organisations struggle to get insight in and address ML systems that are imposing vulnerabilities that can lead to harm for citizens. Therefore, they only get insight if harm is already imposed, while it is better to prevent harm. Furthermore, the non-transparency of organisations raises the question of why they would not be transparent about it, do they have something to hide? As a result, this strengthens the distrust of civil society organisations towards organisations that use ML. Good approaches and efforts to use ML responsibly are thus not seen and acknowledged either.

Fourthly, there are reasons why organisations could be non-transparent about the use and the inner working of ML systems. Being transparent may hamper their competitive position and the possession of intellectual property. Besides that, if people know how an ML model takes decision, they could 'game' the results, In other words, they can adapt their behaviour to get a certain outcome. This may impose risks, for example if the ML system is used to detect financial crime.

7.8 Challenge 7: Interpreting regulations applicable to ML systems

The financial sector is highly regulated, but what it means for the use of ML and what future regulations will require from organisations is not yet crystallised.

There is an extensive legal framework in place in the Netherlands, consisting of among others the European Convention on Human Rights, The Constitution, criminal law, administrative law, sectoral legislation, and the GDPR. The emergence of AI and ML fruits the discussion whether current legislation is satisfactory to cover the risks it entails, or that new legislation is needed. The reflex with a new technology such as ML is to write entirely new legislation, without considering what the current legislation already covers. This is not necessarily the right approach according to the DNB, because current legislation covers already many risks of ML.

In contrast, several civil society organisations consider the current legislation insufficient to provide enough legal protection and human rights protection for citizens (i.e. Platform Bescherming Burgerrechten, Privacy First, Bits of Freedom, and Amnesty International). In April 2021, a proposal for the so-called EU AI act has been published, which has the objectives to ensure the use of AI systems safely that respect existing law on fundamental rights and Union values. Additionally, it has the objective to enhance governance and effective enforcement of fundamental rights law and safety requirements that are applicable to AI systems (European Commission, 2021). Despite these objectives, several interviewed organisations find the proposal still insufficient to protect citizens. The representative of Bits of Freedom calls the legal protection for citizens insufficient in enforcing transparency, accountability and explainability (Personal Communication, January 17, 2022). Furthermore, the representative of Amnesty International calls the AI act a step in the right direction, but the current proposal does not adhere to any of their recommendations for the protection of human rights (Personal Communication, January 25, 2022). Besides the inadequacy in protecting human rights, the proposed AI act does not take the dynamic change dimension into account, as the required conformity assessment for high risk applications has to be performed before deployment of ML systems, but does not cover the dynamic nature of ML systems over time. This is a problem that is challenging for regulatory bodies (Personal Communication representative DNB, January 18, 2022). As summarized by the representative of the Autoriteit Persoonsgegevens (Personal Communication, January 24, 2022): "The objective should be to encourage responsible innovation and protect the fundamental rights of people simultaneously. Yet, there cannot be a strong line identified how those are reconciled in the proposal of the AI act at this moment".

The enforcement of legislations lies in the hands of several regulatory bodies, among which the Autoriteit Persoonsgegevens and the DNB. At the same time, financial institutions have full responsibility to organise internal supervision and compliance with the GDPR. According to the representative of the Autoriteit Persoonsgegevens (Personal Communication, January 24, 2022), organisations have to be very mature to be able to do so. The privacy officer in bank A endorses that the bank has organised its own internal supervision, but that there is little external supervision. This lack of external supervision could lead to undetected risks in the use of ML within the financial sector (Personal Communication, January 17, 2022). Furthermore, the daily supervision of the DNB hardly looks at ML systems at the moment. This will change in the future by the AI act, which gives the DNB a mandate to do so (Personal Communication representative DNB, January 18, 2022).

As illustrated, current legislation and the proposed AI act are considered insufficient to enforce a responsible way of developing and using ML systems. Currently, banks are left free to organise this themselves, but the privacy officer of bank A points out that guidance from external higher authorities on how to deal with ML would be very welcome (Personal Communication, January 17, 2022).

How the ultimate AI act will look like, and how it will affect organisations, only time will tell. The GDPR caused a substantial increase of awareness on data protection and privacy among organisations, so the question is whether the AI act has the same effect on AI within organisations. How the AI act will be adopted in organisations will in turn influence the position regulatory bodies take.

7.9 Conclusion Chapter 7

This section aims to answer the fifth sub-question:

5. What are the main challenges identified in ML practice, seen from a sociotechnical ML system perspective?

Chapter 7 presents the main challenges identified in ML practice, based on the results of an inductive analysis of data from eighteen conducted interviews with stakeholders involved with two ML use cases within two different banks, and representatives of civil society organisations and regulatory bodies. This results in seven challenges identified and described in Sections 7.2 up till 7.8.

Underlying to these challenges are important properties of sociotechnical systems. First, all components as well as their interactions and the dynamics occurring in the sociotechnical ML system

and its environment need to be considered (Dobbe, 2022). If not, the definition of the system boundaries will likely become too narrow, excluding relevant components and associated dynamics with other components. Second, the emergent nature of values such as fairness and safety and vulnerabilities in sociotechnical ML systems are overlooked. This can for example be recognised in the way that is dealt with fairness in the use cases. In the email marketing use case, the exclusion of personal data is the only effort performed to prevent biased outcomes, while developing a fair sociotechnical ML system requires a broader view. In the financial crime detection use case, debiasing is used on the data, which is a technique that tries to solve a sociotechnical problem in a technical solution, approached using a narrow technocentric view (Balayn & Gürses, 2021). The emergent nature of safety requires the stakeholders involved in the ML use case and affected by it to resolve fundamental normative differences through tradeoffs and consensus building (Dobbe, 2022). However, this is given too little attention if not all those stakeholders are being involved. Moreover, vulnerabilities that may be present in sociotechnical ML systems and can lead to harm, often have an emergent nature and will be caused by changes over time (Dobbe, 2022). Therefore, it is vital to keep a grip of the sociotechnical system over time, which is often overlooked in practice, as the focus is on the development of the ML system, and not on the operations.

To address these properties and resolve the challenges presented in this chapter, a more integral approach to specifying, developing and operating ML systems is needed (Dobbe, 2022). To take the first step into a more integral approach, the next chapter presents ten guidelines and a sociotechnical ML lifecycle overview that establish a more comprehensive sociotechnical ML systems perspective in the specification of sociotechnical ML systems. These guidelines contribute to resolving the challenges identified in this chapter and dealing with the dimensions introduced in Chapter 4 and analysed in practice in Chapter 6.

Chapter 8

A sociotechnical guide to ML practice

The final objective of this master thesis is to provide practical guidance to stakeholders in ML practice. Initially, the main purpose is to guide stakeholders that are involved (or should be involved) in the specification of a sociotechnical ML system. To that extent, a set of guidelines to be followed in the specification of sociotechnical ML systems is drawn in this chapter. Sociotechnical specification is proposed as a central activity in any ML use case, being the starting point as well as an activity that allows for refinement throughout the entire sociotechnical ML lifecycle. A guideline is an artefact type that can be designed in design science research, which provides a generalized suggestion about system development (Offermann, Blom, Schönherr, & Bub, 2010). These guidelines provide a starting point for the sociotechnical ML lifecycle, and impact the entire lifecycle as such (Gurzick & Lutters, 2009). Guidelines exist at the boundary between theory and practice, which gives the researcher a means to transfer the developed theory, knowledge, insights in this research in a way that is useful for practice (Gurzick & Lutters, 2009). This chapter starts with explaining the novelty of the guidelines in Section 8.1 and how to use the guidelines in Section 8.2. The ten guidelines in Section 8.3 up till Section 8.12. Further, in Section 8.13, the sociotechnical ML lifecycle is presented, which centres the sociotechnical specification throughout the lifecycle of the sociotechnical ML system. Subsequently, the context-dependence of the guidelines is explained in Section 8.14. Thereafter, the implications for MLOps and Deeploy are discussed in Section 8.15. This chapter results in answering the sixth sub-question:

6. What guidelines can guide ML practice in the sociotechnical specification of sociotechnical ML systems?

8.1 Novelty of the guidelines

ML practices currently lean on a long tradition of engineering and computer science, in which mathematical abstractions are used to display problems and their solutions in technical terms (Dobbe et al., 2021). However, sociotechnical complexity and normative stakes that are present in ML systems that are integrated in a real-life context, cannot be addressed solely using a technical lens, but require a more comprehensive lens (Dobbe et al., 2021). In this research, the aim is to take such a more comprehensive sociotechnical systems lens to guide stakeholders involved in ML use cases. This lens aims to see beyond technical practices, and reframes ML system specification as a multidisciplinary practice (Dobbe et al., 2021). Analysing the introduction of technology into a real-life context as a sociotechnical system is not new, but knows a wide range of findings in the sociotechnical systems' literature (Selbst et al., 2019). However, this way of thinking is not yet widely adopted by ML researchers, policymakers, and ML practitioners. This is for example recognised in the way ML research and policymakers respond to the discriminatory effects ML systems can have. These effects to society are squeezed into the technical, prosing debiasing techniques to solve discrimination (Balayn & Gürses, 2021). Debiasing is the main focus of solving discriminatory effects of ML systems in current policy documents, while the debiasing techniques have limitations. They address bias in mere statistical terms, not accounting for the diverse requirements and needs that different stakeholders in the sociotechnical ML system have regarding the fairness of the system (Balayn & Gürses, 2021).

The guidelines proposed in this chapter answer the need for a sociotechnical ML systems view in ML practice as recognised by several scholars (Sendak et al., 2020; Makarius et al., 2020; Mateescu & Elish, 2019; Green & Chen, 2019b; Dobbe et al., 2021; Balayn & Gürses, 2021). These guidelines are grounded by a knowledge base consisting of different facets in sociotechnical ML systems, to create a comprehensive set of guidelines. For example, insights from AI fairness literature, human-machine interaction, machine dynamics, systems engineering, and sociotechnical specification are used. Moreover, insights from two ML use cases, civil society organisations and regulatory bodies in practice resulted in the identification of seven challenges that are encountered in practice. The guidelines are developed to address these challenges, which makes them highly relevant for ML practice. To summarize, the guidelines form a first step into practically interpreting a wide range of findings in the literature that have important implications for the specification of sociotechnical ML systems (Selbst et al., 2019). Vital is that the guidelines aim to consider the social alongside the technical in any ML use case (Selbst et al., 2019).

8.2 Using the guidelines in practice

This chapter presents ten guidelines that are proposed to stakeholders involved in ML use cases. They provide guidance for specification throughout the sociotechnical ML lifecycle, by accommodating a broader sociotechnical ML systems view. The focus of the guidelines is on the sociotechnical specification, that is the start of the lifecycle as well as a central activity throughout the sociotechnical ML lifecycle, and requires continuous refinement. The connections of the guidelines to the dimensions that form the theoretical framework and to the challenges identified in the application environment are summarised in Table 8.1. The guidelines can be used as a starting point in any ML use case, as they reflect a generalised view. However, ML uses cases are highly context dependent, which should be kept in mind when using the guidelines. Dependent on the context, one should reflect on the applicability of each guideline and directions provided with it.

Table 8.1: Overview Guidelines, addressed Challenges and associated Dimensions

Guideline	**Challenges addressed**	**Associated dimensions**
Guideline 1: Multidisciplinary teams should be established at the beginning of the ML lifecycle	Challenge 1, Challenge 5, Challenge 7	Misspecification, Accountability
Guideline 2: System boundaries should be defined by the multidisciplinary team	Challenge 1	Misspecification, Interpretation, Behaviour
Guideline 3: Enable the identification, addressing, and mitigation of vulnerabilities in the sociotechnical specification	Challenge 2	Misspecification, Machine error, Interpretation, Behaviour, Adaptation, Dynamic change, Downstream impact, Accountability
Guideline 4: An initial specification of the sociotechnical ML system should be formulated before starting the experimental stage of the ML lifecycle	Challenge 1, Challenge 2, Challenge 4, Challenge 7	Misspecification, Accountability, Downstream impact
Guideline 5: Provide feedback channels for different stakeholders during development and operations	Challenge 2, Challenge 3, Challenge 5	Misspecification, Machine error, Interpretation, Behaviour, Adaptation, Dynamic change
Guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical ML system in operation	Challenge 2, Challenge 3, Challenge 4, Challenge 6	Interpretation, Behaviour, Adaptation, Dynamic change, Accountability
Guideline 7: Verify and validate the sociotechnical ML system before operationalizing	Challenge 2, Challenge 3	Machine error, Interpretation, Behaviour, Adaptation
Guideline 8: Establish transparency of the sociotechnical ML system	Challenge 3, Challenge 6, Challenge 7	Interpretation, Behaviour, Downstream impact, Accountability
Guideline 9: Create knowledge and communication between stakeholders in the sociotechnical ML system	Challenge 5	Misspecification, Machine error, Dynamic Change, Accountability
Guideline 10: Establish a safe culture and adequate management within the organisation	Challenges overarching	Accountability

In the subsequent sections, the guidelines are presented, along with proposed directions that ML practitioners can follow to operationalise the guidelines in practice. Additionally, for every guide-

line, the advantages for organisations that follow it are sketched, as well as the challenges the guideline helps to solve in practice and the dimensions that are associated with the guideline. Lastly, an analysis of the guidelines applied to the use cases of this research is performed. Based on this analysis, an example of applying the guidelines to one of the use cases is illustrated for every guideline.

8.3 Guideline 1: Establish a multidisciplinary team at the beginning of the sociotechnical ML lifecycle

To be able to specify and develop a sociotechnical ML system, different perspectives should be established in the project team. Involvement can be divided into three levels:

- Stakeholder roles that could be part of the multidisciplinary team: project manager, problem owner, business manager, data provider, data engineer, data scientist, ML engineer, Operations engineer/MLOps engineer, subject-matter expert, and end-users.
- The second line of defence should be consulted to be informed of the applicable legislations and to detect and mitigate potential risks in the envisioned ML system. To this extent, compliance officers, privacy officers, and risk officers could be consulted.
- A dialogue with external stakeholders could be initiated to be informed about larger societal implications of the envisioned ML system, the relation to human rights and to situate the envisioned system in the legal context. To this extent, civil society organisations and regulatory bodies can be addressed.

The multidisciplinary team should take on the following tasks to start with:

- Top-down decision-making should be avoided in the multidisciplinary teams to allow every perspective to establish in decisions made during the ML lifecycle.
- Distribute and register clear roles and responsibilities among the team members.
- Establish communication channels for the entire ML lifecycle. For example, periodically meetings to discuss decisions and progress, establish ways for different stakeholders to ask questions or address concerns.

Advantages of guideline 1 for organisations
Establishing a multidisciplinary project team provides advantages for organisations. First, a relatively large team can divide the cognitive load of all aspects, challenges, vulnerabilities that are important to consider in the ML lifecycle. Second, involving non-technical or business stakeholders establishes commitment. This prevents that models are developed by the technical stakeholders that are not wished, not needed, or not trusted in the organisation.

Challenges addressed by guideline 1
This guideline addresses challenge 1: Defining the system boundaries for the ML lifecycle, as establishing multiple perspectives helps to avoid too narrow system boundaries. Further, challenge 5: Developing knowledge and communication within sociotechnical ML systems is addressed, as multidisciplinary teams enable communication between different roles with different knowledge and expertise. It also addresses challenge 7: Interpreting regulations applicable to ML systems, because insight from second line of defence and external stakeholders on regulation can directly be applied in the ML lifecycle

Dimensions addressed by guideline 1
This guideline addresses the dimensions *Misspecification* and *Accountability*. Misspecification is addressed as involving many perspectives reduces the possibility for gaps in the ultimate specification of the system and enables considering the larger sociotechnical ML system. Accountability is addressed because the guideline prescribes to distribute and register responsibilities among the team members.

Illustration of the contribution of guideline 1
An example where following guideline 1 could have contributed can be found in the financial crime detection use case. In this use case, the end-users were only involved late in the lifecycle, once the

ML system was already developed. This led to a misspecification of the descriptions of the features used: The end-users had to use these descriptions in their investigations, but were not able to, as the descriptions were specified in a technical language by the data scientists. As a result, a translation document had to be created ad-hoc to make the feature descriptions interpretable for the end-users. Guideline 1 prescribes that a multidisciplinary team should be setup at the beginning of the sociotechnical ML system. Following this guideline, a delegation of end-users would have been part of the team from the start. In that case, end-users would have been directly involved in the specification of the feature descriptions, contributing to understandable descriptions for the end-users.

8.4 Guideline 2: Define the system boundaries as multidisciplinary team

The multidisciplinary team should define the boundaries on what will be part of the specification, design and implementation of the system and what lies outside the scope. To define the system boundaries, the team should start with identifying a clear problem or need for which a solution is required. Once the problem is identified, requirements should be negotiated by the team members and complemented by other stakeholders. Requirements should address both the ML lifecycle process as the ultimate sociotechnical ML system that results from the process.

A few directions to define the system boundaries:

- Using ML should be a thoughtful decision, and should solve a problem or address a need in the organisation.
- Beware of defining the system boundaries too narrow, as this could lead to lacking connection of the ML system to the context it is integrated in.
- The decision-making process, and potential role of a human decision-maker should be within the system boundaries.
- Considering the larger implications of people affected by the sociotechnical ML system should be within the system boundaries.
- Specify requirements for the entire sociotechnical ML system, not just for the ML system within.

Advantages of guideline 2 for organisations
First, carefully defined boundaries of the sociotechnical ML systems will increase efficiency and effectivity in the remainder of the ML lifecycle, as the potential need for ad-hoc specifications later is reduced when the scope and application domain is taken into account from the start. Second, requirements that take larger implications of ML systems into account give data scientist a clear framework for model development and choices to take.

Challenges addressed by guideline 2
Guideline 2 addresses challenge 1: Defining the system boundaries for the ML lifecycle, as it provides guidance to defining the system boundaries.

Dimensions addressed by guideline 2
The dimensions addressed by guideline 2 are *Misspecification*, *Interpretation*, and *Behaviour*. Defining appropriate system boundaries reduces the possibility of misspecification. Further, including the decision-making process within the system boundaries helps to address and solve interpretation vulnerabilities in the ML lifecycle. Last, one should be aware of the behaviour dimension and associated vulnerabilities when designing the sociotechnical ML system within the system boundaries.

Illustration of the contribution of guideline 2
To illustrate the contribution of guideline 2, an example of the financial crime detection use case used. As the end-user was not included in the project team, the decision-making process around the ML system to be developed was given limited attention in the specification of the system. The ML system was developed first, before the interaction with the end-user was discussed, as this was only tested ad-hoc. This resulted in many struggles to integrate the ML system in the decision-making process of the transaction monitoring analysts, as they did not understand the feature

descriptions and struggled to interpret the output of the models in their work. Solutions for these problems were developed afterwards. However, if guideline 2 was followed, the decision-making process had been a vital part of analysis for the specification of the sociotechnical ML system. This would have led to a better specification also adhered to the needs of these stakeholders and an easier implementation of the ML system in the decision-making process.

8.5 Guideline 3: Enable the identification, addressing, and mitigation of vulnerabilities in the sociotechnical specification

The identification of potential vulnerabilities in the envisioned system in an early stage inform the specification of the sociotechnical ML system. Vulnerabilities in the ultimate sociotechnical ML system can in this way be prevented, or if not possible be monitored. To identify vulnerabilities, the following directions can be taken:

- Involve the second line of defence in identifying potential vulnerabilities.
- Involve civil society organisations and regulatory bodies to identify vulnerabilities.
- Be aware that vulnerabilities can emerge in the eight dimensions, as described in Section 4.3.
- Possible (but not exhaustive) vulnerabilities are described in Section 4.5.

Advantages of guideline 3 for organisations
The early identification of potential vulnerabilities allows the project team to translate identified vulnerabilities to concrete requirements, formulated as constraint on the design of the system (Dobbe, 2022). Following these requirements in turn will contribute to a responsible sociotechnical ML system.

Challenges addressed by guideline 3
This guideline addresses challenge 2: Dealing with emergent dimensions in the sociotechnical ML system. An early identification of the potential vulnerabilities help specify how to deal with emergent vulnerabilities that are not yet present in the start of the ML lifecycle, but may emerge. This way, mechanisms to control emergent vulnerabilities such as monitoring and evaluation can be established.

Dimensions addressed by guideline 3
This guideline addresses all dimensions, as it encourages project teams to put effort in identifying vulnerabilities that lie in the different dimensions.

Illustration of the contribution of guideline 3
To illustrate the contribution of guideline 3, an example in the email marketing use case is used. In this use case, awareness and addressing of vulnerabilities that can emerge in the Dynamic change dimension is lacking. Changes in underlying data of the ML system are not always adequately communicated between departments within the bank, neither is monitoring in place to detect for example changes in the data distributions caused by these changes. Following guideline 3, the governance structure of data changes should be matured. The departments should have created a communication channel to communicate changes in data structures, and monitoring metrics could help to identify vulnerabilities of the ML system in operation.

8.6 Guideline 4: Formulate an initial specification of the sociotechnical ML system before starting the experimental stage of the sociotechnical ML lifecycle

Before the experimental phase is started, the project team should have formulated an initial specification of the sociotechnical ML system, which is guided by following the first three guidelines. The initial specification should at least consist of:

- Problem/need that is to be addressed by the sociotechnical ML system.

- Objective of the sociotechnical ML system and whom it is meant to be of service to.
- Team members and their roles and responsibilities in the ML lifecycle.
- A program of requirements for the development process and the resulting sociotechnical ML system.
- Thoughtful selection of what data are used.
- Identification of vulnerabilities and mitigation measures.
- Understanding of sociotechnical context around the ML system.
- Applicable governance processes and organisational guidelines.
- Initial indicative risk assessment and compliance assessment by second line of defence.

Advantages of guideline 4 for organisations
An initial specification guides stakeholders in the subsequent steps of the ML lifecycle, and makes sure the project team has reached agreement on the initial envisioned sociotechnical ML system. Moreover, there has been room for consideration about data selection, vulnerabilities, risks and compliance before the actual development is started. This prevents e.g. developing sociotechnical ML systems that are not addressing a problem, or ML models that are not compliant with applicable regulations and thus cannot be used in practice.

Challenges addressed by guideline 4
First, challenge 1: Defining the system boundaries for the ML lifecycle, is addressed as the system boundaries has to be defined in this initial specification. Second, challenge 2: Dealing with emergent dimensions in the sociotechnical ML system, is addressed as an initial identification of vulnerabilities and mitigation measures is part of the initial specification. Furthermore, challenge 4: Recognising the importance of data in the ML lifecycle, is addressed, as the initial specification includes a thoughtful selection of which data are used. Moreover, the selection should be assessed in the risk assessment and compliance assessment. Lastly, challenge 7: Interpreting regulation applicable to ML systems is accommodated by this guideline, as a compliance assessment should be performed

Dimensions addressed by guideline 4
This guideline addresses the dimensions *Misspecification*, *Downstream impact*, and *Accountability*. Misspecification and Accountability are addressed as indicated in guideline 1. Downstream impact is addressed by a thoughtful selection of what data are used, as the selection of data in the beginning has downstream impact on the remainder of the ML lifecycle.

Illustration of the contribution of guideline 4
The financial crime detection use case contained an initial risk assessment and compliance assessment by second line of defence, which provides an example for other use cases and organisations. Bank A has known use cases in the past in which a complete ML system was developed, without consulting the second line of defence. Once the ML system was ready to put to production, the second line of defence was consulted to assess the system, after which it appeared to be non-compliant or exceeded risk acceptance. Therefore, these ML systems could not be put to production. Currently, the bank requires an initial risk assessment and compliance assessment before a team starts the development phase of the sociotechnical ML lifecycle. This way, risk and compliance considerations become an integral part of the specification of the system.

8.7 Guideline 5: Create feedback channels for different stakeholders throughout the development and operations of the sociotechnical ML system

As developing a sociotechnical ML systems as such is an iterative process, not everything can be specified in the initial specification upfront. Therefore, it is key to provide feedback channels for the team members and other stakeholders during development and operations of the sociotechnical ML system. The feedback channels should meet the following objectives:

- Gather feedback and new insights to improve the initial specification.

- Continuously align the sociotechnical ML system with the needs of stakeholders.
- Design the human-ML system interaction together with ML system designer and end-users to accommodate the needed translation from model to usable output.
- Test and improve the interaction of the ML system with the end-user.

Advantages of guideline 5 for organisations
Feedback channels allow for continuous improvement and alignment of elements within the sociotechnical ML system and its interactions. This contributes to a resulting sociotechnical ML system that meets an adequate specification and accommodates the needs of involved stakeholders, including the end-user.

Challenges addressed by guideline 5
Guideline 5 addresses challenge 2, 3, and 5. First, feedback channels accommodate identifying and identification and addressing vulnerabilities within emergent dimensions during the entire ML lifecycle. As such, guideline 5 addresses challenge 2: Dealing with emergent dimensions in the sociotechnical ML system. Second, Guideline 5 guides challenge 3: Introducing a human decision-maker in an ML decision-making process, as it prescribes designing the ML system together with the end-user and to improve the specification of the decision-making process when needed. Lastly, feedback channels address challenge 5: Developing knowledge and communication within sociotechnical ML systems, as the feedback of different stakeholders adds to the knowledge of system developers, and a form of communication is established by the feedback channels.

Dimensions addressed by guideline 5
The guideline addresses the dimensions *Misspecification*, *Machine error*, *Interpretation*, *Behaviour*, *Adaptation*, and *Dynamic change*. Misspecification and machine error are addressed, because feedback channels provide a means to point out initial misspecification and machine errors, which can then be solved. Interpretation is addressed because the feedback channels can continuously the design of interaction between the ML system and its environment. The dimensions behaviour, adaptation and dynamic change mainly involve emergent vulnerabilities, that are hard to envision upfront. Feedback channels can help to identify emerging vulnerabilities and address these in an improved specification.

Illustration of the contribution of guideline 5
This guideline was well-established in the email marketing use case. Since the project team was quite small, consisting primarily of five persons, the team had regularly meetings to discuss feedback during development of the system and the first period after operationalization. If bank B develops new ML use cases in the future, that require larger project teams and involve more stakeholders, the feedback channels might be specified differently to still be efficient. For example, a technical tool could be used to be able to receive feedback from end-users regarding predictions. Further, periodically feedback sessions need to be planned and managed.

8.8 Guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical ML system in operation

Besides feedback channels, monitoring and evaluation mechanisms should be specified to keep control over the sociotechnical ML system over time and improve it when needed. This is especially important due to the emergent nature of vulnerabilities in sociotechnical ML systems. The following directions can be used:

- For the ML system, performance monitoring can be used as provided by MLOps tools and platforms. Performance monitoring allows for example for monitoring prediction volume, errors occurring, concept drift, and feature distribution.
- For the ML system, MLOps tools such as a version control systems and data version control can be used to ensure reproducibly and traceability of models, experiments, predictions and data sets.
- The decision-making process and decisions of human decision-makers should be periodically assessed.

- Evaluation of the entire sociotechnical ML system should be established, not only for the different subsystems. Auditing and performance assessment could be a way to establish this (Dobbe, 2022).
- Periodical evaluation of the sociotechnical ML system by the project team and potential other stakeholders should be in place.
- Evaluation should include the implications of the sociotechnical ML system for affected people, using a holistic perspective, analysing the negative impact the new system might have on the entire, original environment (Balayn & Gürses, 2021).

Advantages of guideline 6 for organisations Monitoring and evaluation mechanisms allow organisations to keep a grip on the sociotechnical ML system in operation. They help identify emergent vulnerabilities to be solved, concerns of stakeholders that can be addressed and initiate potential improvements.

Challenges addressed by guideline 6
This guideline addresses challenge 2, 3, 4, and 6. Challenge 2: Dealing with emergent dimensions in the sociotechnical ML system is addressed, as monitoring and evaluation mechanisms provide means to detect and solve emergent vulnerabilities. Challenge 3: Introducing a human decision-maker in an ML decision-making process is addresses as the evaluation of the interaction and the assessment of the human decision-makers' work is facilitated by following this guideline. Further, it provides means to keep grip on the influence of data as described Challenge 4: Recognising the importance of data in the ML lifecycle by data version control and performance monitoring. Also, MLOps tools provide transparency described in Challenge 6: Providing transparency of ML systems and outcomes. This is supported by MLOps tools that enable for every prediction to reproduce how it came about.

Dimensions addressed by guideline 6
Guideline 6 addresses the dimensions *Interpretation*, *Behaviour*, *Adaptation*, *Dynamic Change*, and *Accountability*. The assessment of the human decision-maker and evaluation of the interaction addresses the interpretation dimension. The dimensions behaviour, adaptation and dynamic change mainly involve emergent vulnerabilities, that are hard to envision upfront. Monitoring and evaluations can help to identify emerging vulnerabilities and take appropriate measures. Accountability is addressed by the establishment of reproducibility and traceability of models, experiments, predictions and data sets.

Illustration of the contribution of guideline 6
This guideline is illustrated by the email marketing use case. The sociotechnical ML system developed in this use case is periodically evaluated by the project team, approximately two times a year. However, not all relevant aspects of the sociotechnical ML system are part of this evaluation, as it is primarily oriented to evaluate whether the business goals specified are reached. This business-oriented evaluation should be expanded to consider the total sociotechnical ML system, consisting of the functioning of the ML system, the human-ML system interaction, and the impact of the system on customers to prevent and detect harmful impact, that is prescribed by guideline 6. Because the evaluations are too narrow at the moment, relevant emergent vulnerabilities in the ML system and potential negative impact on the organisation and customers of the sociotechnical ML system can be looked. Following guideline 6 contributes to establishing comprehensive evaluations.

8.9 Guideline 7: Verify and validate the sociotechnical ML system before operationalizing

Before operationalizing the sociotechnical ML system in the organisation to become 'business as usual', the system should be verified and validated. This is important because potential issues can be identified and solved before the sociotechnical ML system becomes part of the business. Rather than just the verification and validation that is part of the ML lifecycle, verification and validation of the entire sociotechnical ML system should be performed. The verification of the sociotechnical ML system assesses if the system meets its specification (Dobbe et al., 2021). This can be approached by answering the following questions:

- Does the sociotechnical ML system solve the problem/address the need?

- Does the sociotechnical ML system meet its objective?
- Is adhered to the requirements and constraints?
- Does the sociotechnical ML system have unfair outcomes for (groups of) affected people that should be solved?
- Is the sociotechnical ML system compliant with regulations?

The validation is meant to assess how the ML system performs in its empirical context (Dobbe et al., 2021). To do so, a validation period can be initiated, where the sociotechnical ML system is actually operating in its actual integration context, but the business does not rely on its outcome yet. In such a validation period, the following aspects can be part of validation:

- Is the ML model's functioning and output as expected in the experimental stage?
- Does the ML model operate as expected as part of the ML system and application?
- How does the ML system interact with other technical systems?
- How does the ML system interact with human decision-makers and other stakeholders?
- Do vulnerabilities emerge in the sociotechnical ML system?
- Does the introduction of the sociotechnical ML system have negative impact on the entire, original environment (Balayn & Gürses, 2021)?

Advantages of guideline 7 for organisations Verification and validation allow organisations to confirm whether a developed sociotechnical ML system is ready to become business as usual. Issues can be detected and solved before they can lead to harms, and confidence in the resulting sociotechnical ML system is fostered within the organisation.

Challenges addressed by guideline 7
The guideline addresses challenge 2 and 3. Challenge 2: Dealing with emergent dimensions in the sociotechnical ML system, is addressed as emergent vulnerabilities can be identified in the validation period, before they appear in the real application of the system. Besides that, introducing a human decision-maker, as described in challenge 3: Introducing a human decision-maker in an ML decision-making process, is part of the validation.

Dimensions addressed by guideline 7
The dimensions *Machine Error*, *Interpretation*, *Behaviour*, and *Adaptation* are addressed by guideline 7. Verification and validation allows the project team to identify potential machine error. Further, identification of emergent vulnerabilities in the interpretation, behaviour and adaptation dimensions is accommodated in the validation of the sociotechnical ML system in its actual integration context.

Illustration of the contribution of guideline 7
The contribution of guideline 7 is illustrated by an example in the financial crime detection use case. In this use case, one of the ML models was only validated separately, not being integrated in its application environment. This resulted in a seemingly good model being put to production and becoming business as usual. However, once put to production, it appeared that the ML system produced a lot of incorrect model output, which the transaction monitoring analysts had to use in their work. Guideline 7 prescribes that validation should be performed on the sociotechnical ML system in its actual integration context, before becoming business as usual. If guideline 7 had been followed in the financial crime detection use case, this would have enabled the project team to identify the unexpected incorrect model output once the ML model was integrated in its application environment, before it became business as usual.

8.10 Guideline 8: Establish transparency of the sociotechnical ML system

Transparency is seen by civil society organisations and regulators as essential property of sociotechnical ML systems. At the same time, transparency is often lacking at organisations. Guideline 8 provides directions for organisations to establish transparency of sociotechnical ML systems. The following directions could be used:

- Establish governance for what decision-making processes an ML system or its output are used within the organisation.
- Use data sheets for data sets to document data choices, see Gebru et al. (2018).
- Use model cards for model reporting to document model choices, see Mitchell et al. (2018).
- Use accident investigations, taking a holistic perspective, in case of harm caused by the sociotechnical ML system to draw lessons and potential conclusions to improve the system (Dobbe, 2022).
- Use reporting systems where issues can be reported before they turn into hazards (Dobbe, 2022).
- Use explainability methods to improve interpretability of model output for the organisation, customers, and regulators.
- Provide channels for affected people of ML decision-making to ask questions, express concerns and file complaints.
- Establish transparency on decisions that affect citizens. A citizen should always be able to get an explanation on a decision that impacts him/her.
- Establish an algorithm register for external organisations and society to consult.
- Establish an open dialogue with regulatory bodies and civil society organisations.

Advantages of guideline 8 for organisations
First, transparency and documentation about system development helps to pass on knowledge about the sociotechnical ML system to new employees over time in an organisation. Further, means to investigate accidents and report issues can serve as input to improve the system. Explainability methods can be useful to improve interpretation, but more insight in what methods are appropriate in which situation, and what the effect is on the quality of the decisions is needed before simply use an explainer. Furthermore, transparency to the public fosters trust between organisations and, more importantly, gives citizens about whom a decision is taken possibilities to address these decisions, ask an explanation and express disagreement if needed. Lastly, an algorithm register providing information on ML models and other algorithms that are used in the organisation fosters confidence among the public and make a dialogue with outside organisations possible. This direction should be followed by governmental organisation, but will provoke the arguments of Intellectual Property and competitiveness among commercial organisations. Although it is reasonable that commercial organisations will not make the entire codebase behind the models they use public, they can use some sort of algorithm register. By providing basic information on whether they use ML models or algorithms and for what purpose, the transparency will substantially improve, without sharing Intellectual Property or losing competitive advantage.

Challenges addressed by guideline 8
Guideline 8 addresses challenge 3 and 7. For Challenge 3: Introducing a human decision-maker in an ML decision-making process, using explainability methods can support the human decision-maker and in reporting systems, he can report issues. Challenge 6: Providing transparency of ML systems and outcomes, is addresses as guideline 7 provide several directions to provide transparency of the ML system.

Dimensions addressed by guideline 8
This guideline addresses dimensions *Interpretation*, *Behaviour*, *Downstream Impact*, and *Accountability*. Being able to interpret model output is supported by transparency, for example by explainability methods. Being transparent may have adverse impact on behaviour, as subjects of ML decision-making may adapt their behaviour to game the output. Further, Downstream impact is addressed as governance on the use of ML systems and output should be established to prevent errors in one output to work its way through the organisation without anyone noticing it. Lastly, transparency highly improves the accountability of sociotechnical ML systems.

Illustration of the contribution of guideline 8
This guideline is illustrated by an example of the email marketing use case. In this email marketing, a lack of transparency toward the customers that are affected by the sociotechnical ML

system is identified. First, customers are not specifically noted that they receive marketing emails based on a selection process using an ML system. Second, the stakeholders involved from bank B are not able to explain why a specific customer is selected to receive emails. Following guideline 8 would have established more emphasis on the need to provide transparency of the use of ML to select customers. First, a notification could be added to the emails to inform the customers ML is used to select them. Second, explainability techniques could be used to better be able to explain selections to customers and other stakeholders.

8.11 Guideline 9: Create knowledge and communication between stakeholders in sociotechnical ML systems

Specifying, developing and operating a sociotechnical ML system requires different types of knowledge and expertise to come together. Communication between stakeholders with different expertise is often difficult and lacking within organisations, which negatively influences the effectivity and efficiency of the sociotechnical ML systems. Further, as guideline 1 prescribes to establish multidisciplinary teams, this will increase the tensions between stakeholders. This is the case because stakeholders with different expertises have different, potentially conflicting interests and perceptions of the "reality" (de Bruijn & Herder, 2009). Cooperation as such cannot be taken for granted, but is essential for a well-functioning sociotechnical ML system in the end (de Bruijn & Herder, 2009). To address this, a few directions can be followed:

- Involve technical stakeholders into analysing the broader implications of sociotechnical ML systems to affected people and society. If these stakeholders become aware of potential vulnerabilities and risks, they can take these into account in the development of the ML system.
- Educate business stakeholders up to board level on the basics of ML.
- To establish communication and understanding between technical and non-technical people that speak different 'languages', introduce the role of a linking pin that speaks both 'languages'.
- Establish ML expertise in organisations that want to use ML but do not have the expertise in-house. This way, reliance on external ML development firms can decrease over time.

Advantages of guideline 9 for organisations
Creating knowledge and communication is a challenge, but investing in them will bring efficiency in developing sociotechnical ML systems and deliver better systems.

Challenges addressed by guideline 9
Guideline 9 addresses challenge 5: Developing knowledge and communication within sociotechnical ML systems. This is a great challenge, that will not be solved in one day, but following this guideline provides the first steps.

Dimensions addressed by guideline 9
This guideline addresses the dimensions *Misspecification*, *Machine error*, and *Dynamic change*. Establishing knowledge and communication between different disciplines will lead to better specifications, and thus avoid misspecification. Besides that, the risk of machine error is lowered when technical stakeholders take the vulnerabilities and risks into account when developing the ML system. For example, they become more aware of potential bias that they can cause during model development. Lastly, if an organisation has ML expertise in-house, they can better deal with dynamic change, for example when changing data distribution requires retraining of the ML model, they are not reliant on the availability of an external firm to do so.

Illustration of the contribution of guideline 9
The contribution of guideline 9 is illustrated with the financial crime detection use case. In this use case, a large gap is identified between the "technical" and "non-technical" people. The "non-technical" transaction monitoring analysts ultimately need to work with the ML output, but generally do not understand the basics of ML. On the other hand, the development of the solution is primarily in hands of the "technical" data science team and ML engineering team, that do not have the operational expertise of the transaction monitoring analysts, larger implications of the ML system, compliance, risk, and business goals. This has led to misalignments in the use case

between the disciplines. As this use cases involves many stakeholders, a project manager that functions as a linking pin of the disciplines could benefit in an effective and safe development of the sociotechnical ML system, as prescribed by guideline 9 (and guideline 10).

8.12 Guideline 10: Establish a safe culture and adequate management within the organisation

A safe culture and adequate management in the organisation are vital to be able to establish all other guidelines (Leveson, 2012). A few directions that can be taken:

- Avoid blame cultures by making stakeholders feel safe to share suggestions on what should be approved to the stakeholders who can make improvements (Dekker, 2016).
- Provide a process to share these suggestions separate from performance assessments (Dekker, 2016).
- The board level of an organisation should take responsibility for establishing the culture and incentives to responsibly develop and use ML by creating the rights incentives and sufficient resources for employees (Personal Communication ML engineer bank A, January 24).
- Establishing the proposed guidelines in an organisation should be taken up or supported by management.
- Appoint a project manager that has the role to provide a holistic perspective on the development of the sociotechnical ML system as a whole.

Advantages of guideline 10 for organisations
Establishing a safe culture is a challenge, but is essential if ML is used for increasingly impactful decision-making. It will give people freedom to express concerns, which can be used to improve systems and prevent harms. Further, if management takes up the guidelines, this will help to consolidate them effectively and efficiently in the organisation.

Challenges addressed by guideline 10
Guideline 10 does not directly address a specific challenge. It is rather an overarching guideline to establish the proposed guidelines for specification of sociotechnical ML systems.

Dimensions addressed by guideline 10
This guideline directly addresses *Accountability*, as it contributes to establishing adequate accountability in sociotechnical ML systems by preventing a blame culture. Indirectly, a safe culture and adequate management will contribute to all handling all other dimensions, because the right incentives and awareness increases the sense of importance to tackle the vulnerabilities that arise in each dimension.

Illustration of the contribution of guideline 10
The contribution of guideline 10 is illustrated by the email marketing use case. Bank B in the infancy of using ML in the organisation. This is recognised for example by the lack of internal guidelines, governance structures, specific risk and compliance assessments of sociotechnical ML systems. As the bank continuous to use new sociotechnical ML systems, potentially with higher risk applications and more potential negative impact on customers, the bank needs to mature regarding ML. Therefore, the guideline 10 prescribes that these guidelines should be taken up by management within the organisation. This will contribute to a broad adoption of the guidelines within the organisation, which will help bank B to start new ML use cases in an effective and safe manner.

8.13 Sociotechnical ML lifecycle

Following from the development of the guidelines, the general design of the ML lifecycle illustrated in Section 3.1 needs to be adapted to accommodate the sociotechnical specification throughout the lifecycle of the sociotechnical ML system. As a result, Figure 8.1 presents the sociotechnical ML lifecycle, and the guidelines that relate to the different activities. The arrows present activities that follow up on each other and feedback channels that may lead to changes in the deliverables

of earlier performed stages. As can be seen, the activities in the sociotechnical specification are not linear, which is why they are not connected with arrows. The only requirement is that an initial specification should be formulated before moving to the experimental stage. Besides that, the monitoring and evaluation of the sociotechnical ML system in the operations co-exist, thus also does not represent a linear process. Lastly, guideline 8, 9, and 10 are separately presented as they do not address a specific activity, but are guidelines that should be taken into account throughout the sociotechnical ML lifecycle.

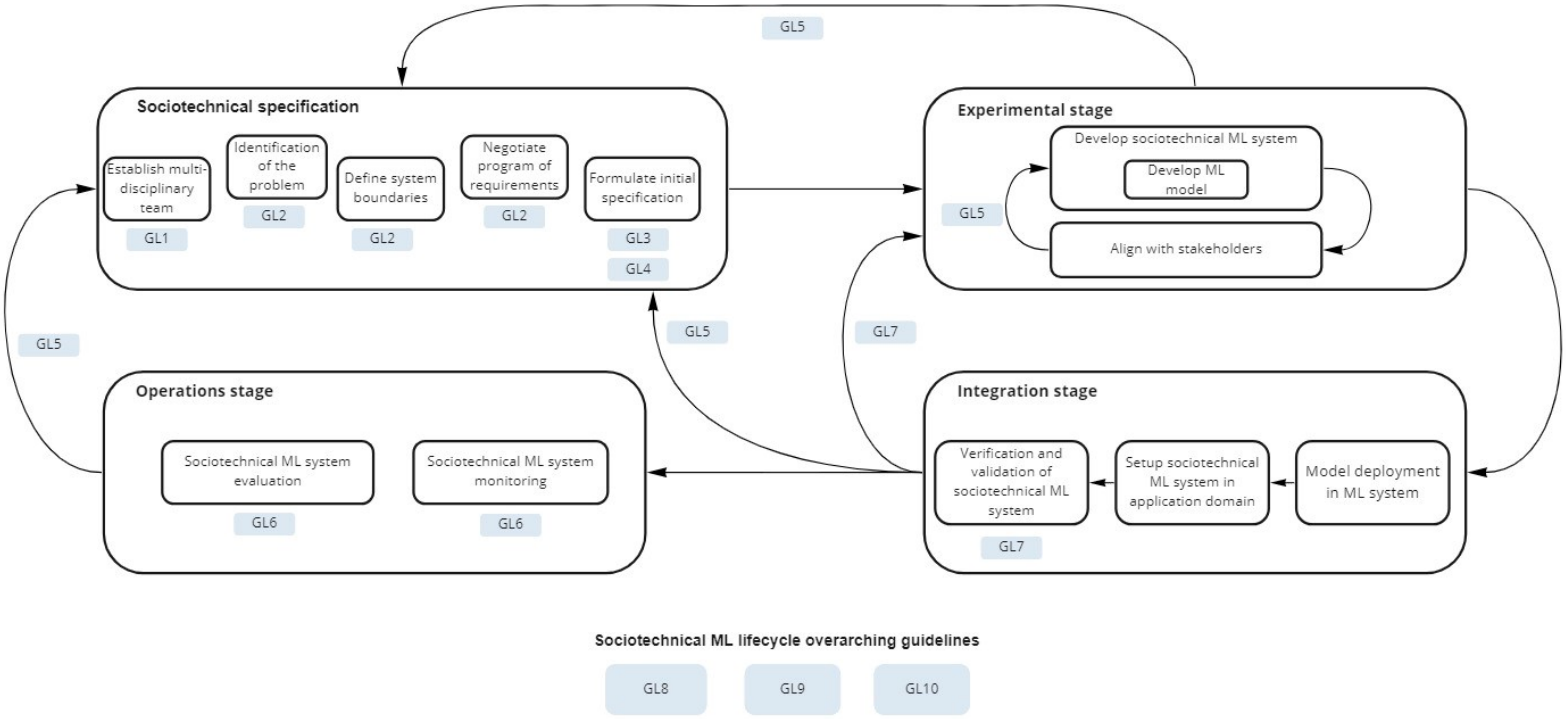

Figure 8.1: Visualisation of the sociotechnical ML lifecycle

8.14 Context-dependence of taking up the guidelines

Taking up the guidelines is expected to be dependent on the context of the ML system and the organisation. There are a few aspects that are extected to influence the adoption of the guidelines by organisations. The first aspect is the objective of the ML system. If the ML system is being developed for a mission-critical decision-making process within the organisation, or has high impact to their customers, the organisation is likely to experience the challenges that the guidelines help to solve. As such, the importance of taking up the guidelines is high, which will encourage stakeholders in the organisation to do so. Second, the maturity of developing and using ML within the organisation is a determining aspect. If the organisation is in its infancy of developing or using ML, it is likely that the emphasis of developing an ML use case is on actually developing a model and getting it in production, which provide large challenges on its own. However, if the way of working regarding ML development is not yet established in those organisations, the guidelines provide a good starting point to actually establish a way of working from scratch. In more mature organisations regarding ML, it is more likely that those have already encountered certain challenges that require a more comprehensive sociotechnical ML system view. Those organisations will probably recognise the challenges described in Chapter 7, which make them aware of the relevance and need of following the guidelines. This will contribute to the guidelines being taken up by the organisation. The last aspect is the dependence of the direction new regulations will take and how the regulatory bodies will respond. If regulations on ML systems becomes stricter and more demanding, the guidelines will become increasingly relevant for all affected organisations.

8.15 Implications of guidelines and sociotechnical ML lifecycle for MLOps practices and Deeploy

The guidelines initially guide the process of sociotechnical ML system specification. Following the guidelines potentially has implications on the technical artefacts and engineering practice in ML use cases. Those implications will merely be context specific. However, this section presents some initial implications the guidelines and the sociotechnical ML lifecycle have on the state-of-the art MLOps practices presented in Chapter 3, which in turn will influence the way technical artefacts such as the ML model and automated pipelines should be developed. Section 3.4 presented a reflection on MLOps practices from a sociotechnical ML systems perspective. Part of the guidelines address the limitations identified in state-of-the-art MLOps practices, and take on the promises of MLOps practices, as described below. Further, the specific implications for the company Deeploy are discussed.

First, in Section 3.4, a limitation of Continuous Integration was identified, as the automated model validation it provides does not consider validation of the larger sociotechnical ML system, but only validated the ML model. This limitation is addressed by guideline 7, that prescribes that validation of not only the ML model, but of the entire sociotechnical ML system is required before operationalizing a sociotechnical ML system or a new model version within the established sociotechnical ML system. Second, it was concluded that human control on deploying new model versions, which is proposed by Deeploy, does not necessarily safeguard the deployment of new model versions, as it can be just clicking a button. This limitation is addressed as well by guideline 7, as verification and validation of the sociotechnical ML system are required before deploying new model versions. As concluded in Section 3.4, MLOps practice offers promising solutions to improve reproducibility, traceability and accountability of ML system. Those solutions are also recommended for monitoring ML systems, as described in guideline 6. Last, MLOps practice solely focuses on the technical functioning and making the ML lifecycle more efficient, as stated in Section 3.4. The guidelines proposed in this chapter aim to widen this technocentric view. The proposed sociotechnical ML lifecycle presented in Figure 8.1 summarizes this.

The guidelines have specific implications of the company Deeploy as well, as the platform contributes to accommodating several guidelines. Moreover, future additions to the platform could be established to expand these contributions. Further, recommendations are given for the larger solution design activities the company could perform besides offering the platform. These insights are presented in D. Table **??** provides an overview of the guidelines, and the relation to Deeploy. In the second column, the table shows whether the guidelines need to be primarily be accomplished by organisational effort, and/or if technical tools may contribute to accomplishing the guideline. The third column shows if the Deeploy platform already provides technical support for each of the guidelines, and if future additions to the platform's functionalities would contribute to the guidelines.

Table 8.2: Overview of guidelines in relation to Deeploy

Guideline	**Implications and/or potential to be supported technically**	**Accommodated by Deeploy platform or could be added in the future**
Guideline 1: Establish a multidisciplinary team at the beginning of the sociotechnical ML lifecycle	Organisational implications, technically supported	No, Future addition
Guideline 2: Define the system boundaries as multidisciplinary team	Organisational implications	No
Guideline 3: Enable the identification, addressing, and mitigation of vulnerabilities in the sociotechnical specification	Organisational implications, technically supported	Yes + Future addition
Guideline 4: Formulate an initial specification of the sociotechnical ML system before starting the experimental stage of the sociotechnical ML lifecycle	Organisational implications	No
Guideline 5: Create feedback channels for different stakeholders throughout the development and operations of the sociotechnical ML system	Organisational implications, technically supported	Yes + Future addition
Guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical ML system in operation	Organisational implication, technically supported	Yes + Future addition
Guideline 7: Verify and validate the sociotechnical ML system before operationalizing	Organisational implication, technically supported	Yes
Guideline 8: Establish transparency of the sociotechnical ML system	Organisational implication, technically supported	Yes + future addition
Guideline 9: Create knowledge and communication between stakeholders in sociotechnical ML systems	Organisational implication	No
Guideline 10: Accommodate a safe culture and establish adequate management within the organisation	Organisational implication	No

8.16 Conclusion Chapter 8

This section aims to answer the sixth sub-question:

6. What guidelines can guide ML practice in the sociotechnical specification of sociotechnical ML systems?

Chapter 8 presents the results from the design phase of this research. This design reflects the gained insights from the knowledge base and the application environment in ten guidelines and a sociotechnical ML lifecycle visualisation. Taking on these guidelines will convey the sociotechnical ML systems perspective into ML practice, by providing directions on how to start an ML use case, centralizing sociotechnical specification throughout the sociotechnical ML lifecycle. The guidelines are presented in Section 8.3 up till Section 8.12. Placing the guidelines in the ML lifecycle requires an adaption of the general ML lifecycle to a proposed sociotechnical ML lifecycle, as presented in Section 8.13.

Following the guidelines contributes to a more effective and safer sociotechnical ML system development and operations, as putting effort in a comprehensive specification of the system upfront and during the sociotechnical ML lifecycle brings several benefits. First, defining the system boundaries and involving a variety of stakeholders contribute to a more complete specification from the beginning. This prevents unnecessary and inefficient ad-hoc specifications during the sociotechnical ML lifecycle. For example, if certain important subsystems or stakeholders are not considered initially and need to be included later, this will lead to a lot of work to be redone. Second, the guidelines require consideration of the implications ML systems can have on people, which contributes to identifying vulnerabilities and mitigate them to prevent harmful systems. Last, the guidelines represent the emergent nature of sociotechnical ML systems. This will guide practitioners to focus more on the specification of the sociotechnical ML system in the operations stage, and how to keep grip on it over time.

It is important to understand that following the guidelines is not straight-forward, and will bring up new challenges. Firstly, integrating a sociotechnical ML systems view means that the social is considered alongside the technical in any ML use case (Selbst et al., 2019). Human behaviour cannot be engineered, but can be unexpected and brings new challenges. For example, a multidisciplinary team should be established as prescribed by the first guideline. Increasing the multidisciplinarity of

the team increases the divergence of their interests and needs, that might be conflicting (de Bruijn & Herder, 2009). Hence, cooperation between them cannot be taken for granted, but is essential to establish the sociotechnical ML system (de Bruijn & Herder, 2009). Therefore, tradeoffs have to be made and consensus needs to be reached, which can be a difficult process (Dobbe et al., 2021). Moreover, including the stakeholders that are affected by the sociotechnical ML system within the system boundaries of analysis will further increase the difference in rights, needs and interests, that might be in conflict (Dobbe, 2022). Further, the guidelines provide directions on how to follow them and establish them into the specification of the sociotechnical ML system, but it is challenging to determine when a guideline is adequately established. Determining this requires applying the guidelines in real-life ML use case settings, to identify best practices and improve the guidelines over time.

Chapter 9

Conclusion and Discussion

This objective of this research was to develop guidelines for specification of sociotechnical Machine Learning (ML) systems, that can be used by practitioners as a starting point for the ML lifecycle. The results presented in the previous chapters are discussed in this chapter, to clarify the meaning of the results and learnings. This chapter presents the main findings, the limitations of the research, contributions of the research to the knowledge base and to practice, and recommendations for future research.

9.1 Main findings: Guidelines for the specification of sociotechnical ML systems

Conducting this research was motivated by Machine Learning being increasingly introduced in decision-making processes in high-stakes domains. To safely and effectively do so, a comprehensive sociotechnical systems view is needed. A described in Chapter 1, two knowledge gaps were identified. First, a knowledge gap was identified in theory on the emergence of vulnerabilities in sociotechnical ML systems. Although it was clear that vulnerabilities do not only entail technical vulnerabilities in ML models, but require a more compressive view to identify them in the larger sociotechnical ML system, a comprehensive overview of vulnerabilities that emerge in sociotechnical ML systems and how to synthesise these in sociotechnical ML system's dimensions, was lacking. This research has investigated these vulnerabilities and constructed a theoretical framework of eights dimensions in which those vulnerabilities occur. The second research gap showed that there was little insight in how was dealt with dimensions of vulnerabilities in the development and use of ML systems in practical settings. This gap is addressed by gaining empirical insights from ML development and use in practice, which resulted in sociotechnical challenges. Ultimately, the gained insights by filling these knowledge gaps resulted in the design of a sociotechnical guide for ML practice. The guide consists of ten guidelines and a proposed adaption of the general ML lifecycle, which presents a sociotechnical ML lifecycle with the guidelines aligned. The sociotechnical guide centres the need for sociotechnical specification, for which the guidelines are developed.

9.1.1 Guidelines for ML practice

The research started with the objective to design guidelines to guide ML practitioners in ML use cases by centralising the specification of sociotechnical ML systems. To capture this objective, the following research question was formulated:

What guidelines should be followed in ML practice to establish a sociotechnical ML systems view in the specification of Machine Learning systems?

A Design Science Research approach was used to ultimately answer the research question. First, a thorough understanding of the scientific knowledge based was reached, resulting in a theoretical framework consisting of eight sociotechnical dimensions in which vulnerabilities emerge. Second, the knowledge gathered served as the basis to analyse ML development and use in a real-world environment, resulting in the identification of seven main challenges, from a sociotechnical ML systems perspective. Ultimately, the insights from the knowledge base and application environment served as input to the design cycle, in which ten guidelines and a sociotechnical ML lifecycle were developed. Table 8.1 presented the ten guidelines, the challenges they address and the dimensions

that are associated to them. The results show that the guidelines provide directions on how to deal with the all the main challenges identified. Further, the guidelines are associated to the different dimensions. The guidelines help to identify and address vulnerabilities that arise in the dimensions.

Following the guidelines contributes to a more effective and safer sociotechnical ML system development and operations, as putting effort in a comprehensive specification of the system upfront and during the sociotechnical ML lifecycle brings several benefits. For example, taking on the guidelines contributes to a more complete specification from in the beginning, preventing unnecessary and inefficient ad-hoc specifications during the sociotechnical ML lifecycle. Second, the guidelines contribute to consideration of the implications ML systems can have on people, which contributes to identifying vulnerabilities and mitigate them to prevent harmful systems. Last, the guidelines represent the emergent nature of sociotechnical ML systems. This will guide practitioners to focus more on the specification of the sociotechnical ML system in the operations stage, and how to keep grip on it over time.

Following the guidelines is not straightforward, and will bring up new challenges. For example, increasing the multidisciplinary of the project team increases the divergence of their interests and needs, that might be conflicting (de Bruijn & Herder, 2009). Therefore, tradeoffs have to be made and consensus needs to be reached, which can be a difficult process (Dobbe et al., 2021). Moreover, including the stakeholders that are affected by the sociotechnical ML system within the system boundaries of analysis, this will further increase the difference in rights, needs and interests, that might be in conflict (Dobbe, 2022). Lastly, it is challenging to determine when a guideline is adequately established. Determining this requires applying the guidelines in real-life ML use case settings, to identify best practices and improve the guidelines over time.

9.1.2 Generalizability of results

In the research, insights from an application environment contributed to the design of the guidelines. This raises the question of how generalizable the guidelines are to other ML use cases. Generalizability is supported in several ways. First, the knowledge base that is laid in this research is based on a general approach on identifying vulnerabilities in sociotechnical ML systems, instead of directly diving into sector or use case specific vulnerabilities. Second, the application environment focussed on two use cases, rather than only one use case. This leads to a richer understanding of how is dealt with the sociotechnical dimensions in practice, as well as which challenges can be identified. Third, the use cases are not the only empirical insights that contributed to the sociotechnical guide. Seven interviews with external civil society organisations and regulators complement the insights on dimensions and relating vulnerabilities in the financial sector and in general. The result is a set of ten guidelines that are generally applicable for ML use cases aimed to be integrated in decision-making processes.

Despite the generally applicable developed guidelines, every use case, every context and every ML system is different. Therefore, using the guidelines in practice will require different suitable implementations. For example, a multidisciplinary team is recommended for every use case, but the size of the team, the persons that will be part of it and the roles they have are depended on the specific situated context.

9.2 Limitations

The limitations of this research are discussed next, as this helps to understand the value of the research results and provide directions for further research.

9.2.1 Design Science Research process

A Design Science Research approach has been used to conduct this research. The main purpose of conducting DSR is to design an artefact. Because there was a lack of theoretical and empirical understanding, the choice was made to first thoroughly conduct research in these areas before moving to the design cycle. This design cycle should iterate rapidly between construction of an artefact, and evaluating it, leading to refinements in the design (Hevner, 2007). Given the time limits, the artefact was not evaluated during this research.

A second limitation in the research process evolves around the selection of the use cases. The use cases entail the development and use of two very different sociotechnical ML systems within two different banks, which enriches the results. However, both use cases were developed for applications in banks, within the financial sector. As developing ML systems is highly context-dependent, choosing a different sector as application environment could have led to other insights on the dimensions, challenges and final guidelines. To complement the insights within the two banks, interviews have been conducted with civil society organisations and regulatory bodies. Those interviews were not entirely focused on ML in banks, but complemented the insights with more general perspectives toward the development and use of ML systems. Nevertheless, it would be valuable to conduct further research into sociotechnical ML systems development and use within other banks, financial institutions and even other sectors.

9.2.2 Theoretical framework

The theoretical framework is constructed bottom-up by means of a literature review, resulting in 24 sociotechnical vulnerabilities that can emerge in sociotechnical ML systems. Therefore, the theoretical framework relies on the thoroughness of the literature review performed. Although the literature review presents many vulnerabilities, it cannot be guaranteed that all important vulnerabilities have been identified. Nevertheless, the construction of the theoretical framework consisting of eight dimensions, provides a basis to identify more relating vulnerabilities, top-down. Further, the vulnerabilities have been generally identified, which provides generalizability of the resulted theoretical framework. However, sociotechnical ML systems specification is highly context-dependent, which means that not every vulnerability will emerge in every context, and that context-dependent vulnerabilities may arise that are not captured in this research.

9.2.3 Sociotechnical ML systems view

In this research, the sociotechnical ML system that is the centre of analysis is bounded to the decision-making process an ML system becomes part of, and the view is used to ultimately guide stakeholders involved in the development and use of ML systems, centring the sociotechnical specification. However, this view does not consider even broader sociotechnical implications within and outside organisations when introducing ML systems. These implications could entail other changes in organisations, institutions and workflows as a result of introducing an ML system. An example could be that the introduction of an ML system causes the organisation to fire half of the employees in a department, which could impose other vulnerabilities. Analysing those implications would require to centre the organisation and the people, instead of the ML system's interaction with social subsystems, as in this work. The choice of the sociotechnical ML systems view and its influence on the theoretical framework have two reasons. The first is that this research's scope has consciously not included other organisational changes and impact on secondary decision-making systems to put focus in the work given the limited time. The second reason is that the literature that could be found on vulnerabilities in sociotechnical ML systems does not consider larger implications on organisations either, which steered the research in the followed direction. As a result, this research provided a first step in widening the technocentric view towards a sociotechnical ML system view, and this line should be continued by widening the sociotechnical ML systems view even more.

9.3 Research contributions and future research

This section elaborates on the contributions this research has to the scientific knowledge base as well as to practice.

9.3.1 Scientific contributions

The scientific relevance of this research is seen in particular in two contributions. First, the scientific relevance is a logical consequence of taking the research steps to address the first knowledge gap: A comprehensive overview of what vulnerabilities can emerge in sociotechnical ML systems is lacking, as well as theory that synthesises in which sociotechnical ML system's dimensions they emerge. The comprehensive overview has been created by the identification of 24 vulnerabilities, as described in Chapter 4. Next, the lacking theory that grasps the vulnerabilities has been developed, by synthesizing the vulnerabilities in eight dimensions: Misspecification, Machine error,

Interpretation, Behaviour, Adaptation, Dynamic Change, Downstream impact, and Accountability. Filling this gap builds upon the work of Dobbe et al. (2021), which stated that taking a sociotechnical ML systems will explain how vulnerabilities originate. This explanation is found in this research in the eight dimensions.

Second, this research contributions to the scientific knowledge base by using established knowledge and combine it in a new context, This is reflected in the bottom-up approach, that starts outlining the vulnerabilities. These vulnerabilities are not only drawn upon specific literature on ML systems, but are enriched with literature on the introduction of other types of machines in sociotechnical dynamic contexts, such as (Rasmussen, 2000), Leveson (2012), (Nickerson, 1998), (Tsymbal, 2004), (Langdon, 1980), and (Goodwin & Fildes, 1999). Integrating insights from this literature to sociotechnical ML systems provides the potential for ML researchers to learn from these insights and build further research on.

Third, basing the empirical part of the research on a solid knowledge base leads to another contribution to the scientific knowledge base. It provides empirical proof of how is dealt with dimensions of vulnerabilities in practice, which is translated to main sociotechnical challenges in practice. These challenges can fuel scientific research, addressing the actual challenges identified in practice.

9.3.2 Practical contributions

The main contribution of this research is the design of a sociotechnical guide, consisting of practical guidelines and a general sociotechnical ML lifecycle aligned with these guidelines. The practical contributions of this sociotechnical guide are divided among three different audiences of this research: ML practitioners, civil society organisations and regulators, and technology providers, in this case aimed to Deeploy, in collaboration with whom this thesis came into existence.

First, the practical contributions to ML practitioners. ML practitioners are organisations, departments within organisations, and individuals, involved in ML use cases or about to start with ML use cases. The sociotechnical guide provides a starting point for ML practitioners to start an ML use case with, and provide points to evaluate existing sociotechnical ML systems on. Using the sociotechnical guide will consolidate a broader sociotechnical ML systems view into ML practice. This will contribute to the design and implementations of sociotechnical ML systems that are safe, effective, and lasting.

Second, the insights from this research can serve civil society organisations and regulators in two ways. The capturing of vulnerabilities they see in ML development and use in society into eight dimensions can help them articulate problems to policymakers and ML system developers. Second, the sociotechnical guide aims to involve these organisations more in the specification of sociotechnical ML systems, whereas currently they mainly come into the picture when ML systems are already used in operation and lead to harms for society. The sociotechnical guide helps them to anticipate on vulnerabilities, instead of intervene when the damage is already done.

Lastly, the research provides practical contributions to Deeploy and other technology providers, as it provides insight into how technological tools can be situated in the sociotechnical ML system, and also which elements require other approaches. The Deeploy platform partly supports implementing guideline 3, 5, 6, 7, and 8. The Deeploy platform support guideline 3, as the platform can support both detecting vulnerabilities that emerge in the operations stage of the sociotechnical Ml system and support mitigating certain vulnerabilities. For example, data shifts and concept drift can be monitored in the Platform. Guideline 5 is supported as the platform provides a feedback channel for end-users to evaluate predictions. Further, the Deeploy platform contributes to guideline 6, as several monitoring mechanisms on the ML system within the sociotechnical ML system are provided. Besides that, Deeploy facilitates deployment and streamlines integration of the ML model in its application environment, which makes it possible to validate the ML system in its empirical context as prescribed by guideline 7. Lastly, the Deeploy platform contributes to transparency by explainability, reproducibility and traceability, contributing to guideline 8.
Moreover, potential future additions are recommended to implement in the Deeploy platform, to further contribute to establishing the guidelines. The first potential addition is to implement governance constraints on certain actions in the platform, such as new model deployment, to keep grip on the ML system. Second, the feedback mechanisms can be expanded by providing different feedback possibilities for different stakeholders, such as business experts, data scientists, compli-

ance and risk officers, and people about whom the ML system predicts something. Lastly, the alert functionality is recommended to be expanded to define other responsible stakeholders to act upon certain alerts besides the data scientist.
Besides the guidelines that Deeploy could accommodate, the research provides insight in the limitations of solving sociotechnical problems with technical solutions, of which Deeploy should be cautious when operating ML systems. Therefore, it is recommended to take the sociotechnical ML systems view before using Deeploy to deploy ML systems, and provide additional support to organisations by establishing the guidelines that require more organisational effort.

9.4 Recommendations for future research

Several recommendations for future research are given in this section.

First, evaluation and demonstration of the use of guidelines in a real-life ML use case setting is recommended. As explained in the limitations, the iterative steps in the design cycle between design and evaluation have not been performed due to time restrictions. A next research step could be to perform these steps, in order to evaluate the guidelines. To do so, researchers could be involved in a real-life setting in which an ML use case is initiated, and the guidelines are used as a starting point. This way, it could be researched what the value of the guidelines is in practice, and provide insights to improve the sociotechnical guide. Also, insights on best-practices of how to go from guideline to implementation can be gathered this way. For example, guideline 6 prescribes that the sociotechnical ML system should be evaluated as a whole. Additional research on how this evaluation can be performed in practice would be beneficial.

Second, this research used the financial sector, with two ML uses cases within banks, as empirical environment to the research. Involving more use cases would contribute to an even richer understanding of how the dimensions consolidate in practice and what challenges are present in practice. Including use cases from other financial institutions would provide insights whether differences between banks and for example insurers could be detected that could inform other guidelines. Researching use cases in other sectors would also enrich the research output. As the financial sector is highly regulated and risk-averse, this reflects in how ML use cases are approached. Including less regulated spaces could lead to different results that complement this research.

Third, the guidelines are initially developed to guide the specification process of sociotechnical ML systems. However, introducing a sociotechnical ML system potentially requires changes in the technical artefacts and ML development and engineering practices as well. It is recommended to take on the sociotechnical ML system perspective to research this, to complement this master thesis by further initializing the sociotechnical ML system perspective into ML use cases.

Lastly, more research into the societal and organisational changes that are caused by the introduction of ML systems is proposed. Instead of centring the ML system, which was the approach in this research, centring the organisation will provide more insight into implications of ML systems on organisations and people. This research widened the technical views to a sociotechnical view, but this view should be widened even more in future research.

References

AFM. (n.d.). *Wet ter voorkoming van witwassen en financieren van terrorisme (Wwft) — Onderwerpinformatie van de AFM — AFM Professionals.* Retrieved from https://www.afm.nl/nl-nl/professionals/onderwerpen/wwft-wet

AI HLEG. (2019). *HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES Definition developed for the purpose of the AI HLEG's deliverables* (Tech. Rep.). Retrieved from https://ec.europa.eu/digital-single-

Alla, S., & Adari, S. K. (2021). *Beginning MLOps with MLFlow.* Apress. doi: 10.1007/978-1-4842-6549-9

Alter, S. (2010). BRIDGING THE CHASM BETWEEN SOCIOTECHNICAL AND TECHNICAL VIEWS OF SYSTEMS IN ORGANIZATIONS. In *International conference of information systems* (Vol. 54, pp. 1–23). Retrieved from http://aisel.aisnet.org/icis2010_submissions/54

Altexsoft. (2021, 7). *What is API: Definition, Specifications, Types, Documentation — AltexSoft.* Retrieved from https://www.altexsoft.com/blog/engineering/what-is-api-definition-types-specifications-documentation/

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020, 9). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society*, *35*(3), 611–623. doi: 10.1007/S00146-019-00931-W

Balayn, A., & Gürses, S. (2021). *Beyond Debiasing Regulating AI and its inequalities* (Tech. Rep.). Retrieved from https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17. doi: 10.1016/j.intcom.2010.07.003

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... Eckersley, P. (2020). Explainable Machine Learning in Deployment. Retrieved from https://doi.org/10.1145/3351095.3375624 doi: 10.1145/3351095.3375624

Brom, D. (2021). *AI Governance in the City of Amsterdam: Scrutinising Vulnerabilities of Public Sector AI Systems* (Tech. Rep.). TU Delft. Retrieved from https://repository.tudelft.nl/islandora/object/uuid%3Abd37fe4c-4c55-4e6f-8b94-9d9a0cdf7dc4

Curtis, S. (2020, 4). *Endpoint vs. API. What is what? — by Steven Curtis — Medium.* Retrieved from https://stevenpcurtis.medium.com/endpoint-vs-api-ee96a91e88ca

de Bruijn, H., & Herder, P. M. (2009). System and actor perspectives on sociotechnical systems. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *39*(5), 981–992. doi: 10.1109/TSMCA.2009.2025452

Deeploy. (n.d.). *Product.* Retrieved from https://www.deeploy.ml/product/

Dekker, S. (2016). *Just Culture : Balancing Safety and Accountability.* CRC Press. Retrieved from https://www-taylorfrancis-com.tudelft.idm.oclc.org/books/mono/10.4324/9781315251271/culture-sidney-dekker doi: 10.4324/9781315251271

Dobbe, Krendl Gilbert, & Mintz. (2021, 11). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*. doi: 10.1016/j.artint.2021.103555

Dobbe, R. (2022). System Safety and Artificial Intelligence. In *Oxford handbook on ai governance.* Retrieved from https://arxiv.org/abs/2202.09292

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018, 7). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. Retrieved from https://arxiv.org/abs/1807.00553v2

DVC. (n.d.). *Versioning Data and Models — Data Version Control · DVC.* Retrieved from https://dvc.org/doc/use-cases/versioning-data-and-model-files

European Commission. (2018). *ARTICLE 29 DATA PROTECTION WORKING PARTY Guidelines on Automated individual decision-making and Profiling* (Tech. Rep.). Retrieved from

https://ec.europa.eu/newsroom/article29/items/612053/en
European Commission. (2021, 4). *EUR-Lex - 52021PC0206 - EN - EUR-Lex.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
Financial Intelligence Unit. (n.d.). *Banken — FIU-Nederland.* Retrieved from https://www.fiu-nederland.nl/nl/meldergroep/8
Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). doi: 10.1162/99608F92.8CD550D1
Friese, S., Soratto, J., & Pires, D. (2018). *Carrying out a computer-aided thematic content analysis with ATLAS.ti.* Göttingen. Retrieved from www.mmg.mpg.de/workingpapers
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2018, 3). Datasheets for Datasets. *Communications of the ACM*, *64*(12), 86–92. Retrieved from https://arxiv.org/abs/1803.09010v8 doi: 10.48550/arxiv.1803.09010
Goodwin, P., & Fildes, R. (1999). Judgmental Forecasts of Time Series Aected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? *Journal of Behavioral Decision Making*, *12*, 37–53. doi: 10.1002/(SICI)1099-0771(199903)12:1
Green, B., & Chen, Y. (2019a, 1). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 90–99. doi: 10.1145/3287560.3287563
Green, B., & Chen, Y. (2019b, 11). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). doi: 10.1145/3359152
Green, B., & Chen, Y. (2020). Algorithm-in-the-Loop Decision Making. In *Proceedings of the aaai conference on artificial intelligence* (pp. 13663–13664). doi: https://doi.org/10.1609/aaai.v34i09.7115
Grønsund, T., & Aanestad, M. (2020, 6). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, *29*(2), 101614. doi: 10.1016/J.JSIS.2020.101614
Gurzick, D., & Lutters, W. G. (2009). Towards a Design Theory for Online Communities. In *Desrist acm.* Malvern.
Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92.
High Level Expert Group on Artificial Intelligence EU. (2020). *Home page - ALTAI.* Retrieved from https://altai.insight-centre.org/
Hove, S. E., & Anda, B. (2005). Experiences from conducting semi-structured interviews in empirical software engineering research. *Proceedings - International Software Metrics Symposium*, *2005*, 10–23. doi: 10.1109/METRICS.2005.24
Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2020). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*. doi: 10.1177/0894439320980118
Johannesson, P., & Perjons, E. (2014). *An Introduction to Design Science.* Springer International Publishing Switzerland. doi: 10.1007/978-3-319-10632-8
Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016, 7). *Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making.* Retrieved from https://hbr.org/2016/10/noise
Kohli, N., Barreto, R., & Kroll, J. A. (2018). Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems. In *1st conference on fairness, accountability, and transparance.*
Koops, B.-J. (2021). The concept of function creep. *Law, Innovation and Technology*, *13*(1), 29–56. doi: 10.1080/17579961.2021.1898299
Kuwajima, H., Yasuoka, H., & Nakae, T. (2020, 5). Engineering problems in machine learning systems. *Machine Learning*, *109*(5), 1103–1126. doi: https://doi.org/10.1007/s10994-020-05872-w
Langdon, W. (1980). Do Artifacts Have Politics? *Daedalus*, *109*, 121–136. Retrieved from http://www.jstor.org/stable/20024652?origin=JSTOR-pdf
Leveson, N. (2012). *Engineering a Safer World.* MIT Press. Retrieved from https://ebookcentral-proquest-com.tudelft.idm.oclc.org/lib/delft/reader.action?docID=3339365
Li, P., Li, T., Ye, H., Li, J., Chen, X., & Xiang, Y. (2018, 10). Privacy-preserving machine learning with multiple data providers. *Future Generation Computer Systems*, *87*, 341–350. doi: 10.1016/J.FUTURE.2018.04.076

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020, 11). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. doi: 10.1016/j.jbusres.2020.07.045

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. doi: https://doi.org/10.1016/0167-9236(94)00041-2

Mateescu, A., & Elish, M. C. (2019). *AI in Context* (Tech. Rep.).

Miller, T. (2019, 2). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. doi: 10.1016/J.ARTINT.2018.07.007

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2018). Model Cards for Model Reporting. In *Fat* 2019 - proceedings of the 2019 conference on fairness, accountability, and transparency* (pp. 220–229). doi: 10.1145/3287560.3287596

Mittelstadt, B. (2019, 6). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. doi: 10.1038/s42256-019-0114-4

MLFlow. (n.d.). *MLflow Tracking — MLflow 1.21.0 documentation.* Retrieved from `https://mlflow.org/docs/latest/tracking.html`

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220.

Offermann, P., Blom, S., Schönherr, M., & Bub, U. (2010). Artifact Types in Information Systems Design Science – A Literature Review. In *International conference on design science research in information systems 2010* (Vol. 6105 LNCS, pp. 77–92). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-13335-0{_}6

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2020, 11). Challenges in Deploying Machine Learning: a Survey of Case Studies.
doi: https://doi.org/10.48550/arXiv.2011.09926

Parasuraman, R., & Manzey, D. H. (2010, 6). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. doi: 10.1177/0018720810376055

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007, 12). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. doi: 10.2753/MIS0742-1222240302

Qu, S. Q., & Dumay, J. (2011). The qualitative research interview. *Qualitative Research in Accounting & Management*, *8*(3), 238–264. Retrieved from `www.emeraldinsight.com/1176-6093.htm` doi: 10.1108/11766091111162070

Rajkomar, A., Dean, J., & Kohane, I. (2019, 4). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. doi: 10.1056/nejmra1814259

Rasmussen, J. (2000). Designing to support adaptation. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 554–557).

Rudin, C. (2019, 5). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. doi: 10.1038/S42256-019-0048-X

Ruf, P., Madan, M., Reich, C., & Ould-Abdeslam, D. (2021, 10). Demystifying mlops and presenting a recipe for the selection of open-source tools. *Applied Sciences (Switzerland)*, *11*(19). doi: 10.3390/app11198861

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). *"Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI* (Tech. Rep.). Retrieved from `https://doi.org/10.1145/3411764.3445518` doi: 10.1145/3411764.3445518

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *NIPS*, 2494–2502.

Selbst, A. D., Friedler, S. A., Venkatasubramanian, S., Vertesi, J., Boyd, D., & Venkatasubrama, S. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Fat* '19: Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68). doi: 10.1145/3287560.3287598

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., ... O'Brien, C. (2020, 1). "The human body is a black box": Supporting clinical decision-making with deep learning. In *Fat* 2020 - proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 99–109). Association for Computing Machinery, Inc. doi: 10.1145/3351095.3372827

Snyder, H. (2019, 11). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. doi: 10.1016/J.JBUSRES.2019.07.039

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the

Machine Learning Life Cycle; A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2021*. doi: 10.1145/3465416.3483305

Syam, N., & Sharma, A. (2018, 2). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Industrial Marketing Management*, *69*, 135–146. doi: 10.1016/j.indmarman.2017.12.019

Torraco, R. J. (2002). Research Methods for Theory Building in Applied Disciplines: A Comparative Analysis. In *Advances in developing human resources* (Vol. 4, pp. 355–376). Sage Publications.

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. doi: 10.1177/1534484305278283

Treveil, M. (2020). *Introducing MLOps: How to Scale Machine Learning in the Enterprise*. O'Reilly Media, Inc. Retrieved from `www.dataiku.com`

Tsymbal, A. (2004, 4). *The Problem of Concept Drift: Definitions and Related Work.* Dublin.

Unger, R. M. (1983). The Critical Legal Studies Movement. *Harvard Law Review*, *96*(3), 561–675. Retrieved from `https://about.jstor.org/terms`

Valohai. (n.d.). *What is a Machine Learning Pipeline?* Retrieved from `https://valohai.com/machine-learning-pipeline/`

van der Burgt, J. (2019). *General principles for the use of Artificial Intelligence in the financial sector* (Tech. Rep.). Retrieved from `https://www.dnb.nl/media/voffsric/general-principles-for-the-use-of-artificial-intelligence-in-the-financial-sector.pdf`

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*. doi: 10.1177/2053951717743530

Veale, M., & Brass, I. (2019). Administration by Algorithm? Public Management meets Public Sector Machine Learning. In *Algorithmic regulation.* Retrieved from `https://ssrn.com/abstract=3375391`

Veale, M., Kleek, M. V., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. doi: 10.1145/3173574.3174014

Visual Studio Code. (n.d.). *Snippets in Visual Studio Code.* Retrieved from `https://code.visualstudio.com/docs/editor/userdefinedsnippets`

Vogelsang, A., & Borg, M. (2019). Requirements Engineering for Machine Learning: Perspectives from Data Scientists.
doi: 10.6084/m9.figshare.8067593.v1

Winter, R. (2008, 11). Design science research in Europe. *European Journal of Information Systems*, *17*(5), 470–475. doi: 10.1057/EJIS.2008.44/FIGURES/3

Zejnilović, L., Lavado, S., Soares, C., Martínez De Rituero de Troya, , Bell, A., & Ghani, R. (2021). *Machine Learning Informed Decision-Making with Interpreted Model's Outputs: A Field Intervention.*

Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence* (Tech. Rep.).

Zhao, Y. (2021). *Machine Learning in Production: A Literature Review.* Retrieved from `https://staff.fnwi.uva.nl/a.s.z.belloum/LiteratureStudies/Reports/2021-LiteratureStudy-report-Yizhen.pdf`

Zhou, Y., Yu, Y., & Ding, B. (2020, 10). Towards MLOps: A Case Study of ML Pipeline Platform. *Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2020*, 494–500. doi: 10.1109/ICAICE51518.2020.00102

Appendix A

Selected literature for integrative literature review

The table below presents the selected literature for the integrative literature review that serves as the basis for Chapter 4, and the main theme of every paper.

Table A.1: Overview selected literature

Title	Reference	Main theme
AI Governance in the City of Amsterdam: Scrutinising Vulnerabilities of Public Sector AI Systems	Brom (2021)	AI governance
System Safety and Artificial Intelligence	Dobbe (2022)	AI safety
"Everyone wants to do the model work, not the data work":Data Cascades in High-Stakes AI	Sambasivan et al. (2021)	Data in ML
Explanation in artificial intelligence: Insights from the social sciences	Miller (2019)	Explainability of AI systems
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead	Rudin (2019)	Explainability of AI systems
Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government	Janssen et al. (2020)	Explainability of AI systems
Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making	Veale et al. (2018)	Fairness and Accountability of AI systems
Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems	Kohli et al. (2018)	Fairness and Accountability of AI systems
Beyond Debiasing: Regulating AI and its inequalities	Balayn and Gürses (2021)	Fairness of AI systems
A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics	Dobbe et al. (2018)	Fairness of AI systems
A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle	Suresh and Guttag (2021)	Fairness of AI systems
Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data	Veale and Binns (2017)	Fairness of AI systems
Fairness and Abstraction in Sociotechnical Systems	Selbst et al. (2019)	Fairness of AI systems
Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making	Kahneman et al. (2016)	Human decision-making
Complacency and bias in human use of automation: An attentional integration	Parasuraman and Manzey (2010)	Human-Machine interaction
Confirmation Bias: A Ubiquitous Phenomenon in Many Guises	Nickerson (1998)	Human-Machine interaction
Judgemental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy	Goodwin and Fildes (1999)	Human-Machine interaction
Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments	Green and Chen (2019a)	Human-ML system interaction
Machine Learning Informed Decision-Making with Interpreted Model's Outputs: A Field Intervention	Zejnilović et al. (2021)	Human-ML system interaction
The Principles and Limits of Algorithm-in-the-Loop Decision Making	Green and Chen (2020)	Human-ML system interaction
The Problem of Concept Drift: Definitions and Related Work	Tsymbal (2004)	Machine dynamics
Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools	Ruf et al. (2021)	MLOps
Engineering a Safer World	Leveson (2012)	Systems engineering
Engineering problems in machine learning systems	Kuwajima et al. (2020)	Systems engineering

Appendix B

Interview protocol for stakeholders involved in use cases

Below, the interview protocol used for the interviews with the stakeholders involved in the ML use cases is presented. The questions are grouped in themes. In each interview, every theme is touched upon, but the asked questions differ from interview to interview.

B.1 Interview questions

Introduction

- Can you describe a use case you have worked on that we can talk about throughout the interview?
- Can you describe your role and responsibilities in the ML use case?
- Which other roles were involved in the ML use case?

Specification

- Can you tell me what the start of the project looked like?
- Why is the model developed?
- What is the goal of the ML model?
- Before the development of the model is started, were there requirements or criteria specified?
- What do requirements look like?
  - Can you give some examples of requirements that were defined?
- How did you arrive at the final set of requirements?
- Which people with which roles are involved in specifying the use case and requirements?
- Were the requirements documented?
- What is done with the requirements throughout the ML lifecycle?
- How is evaluated whether the ultimate use of the ML system adheres to the requirements?

Sociotechnical dimensions and vulnerabilities
Misspecification

- What is considered inside and outside the scope of the use case project?
- How do you deal with the social context the ML system will be integrated in?

Machine error

- What would you consider mistakes/errors that could be made throughout the ML lifecycle?
- What would you consider mistakes/errors that could happen in the use of the model?
- How do you deal with potential errors or biases in the model output?
- What could be the impact of errors and who is impacted?
- What biases could occur and do you deal with biases?

Interpretation

- Would you consider the model developed a black-box or is there some level of interpretability?
- Would you say the model is explainable and why?
- To whom is the model explainable?
- How is the model output used by the user?
- How is the model output presented to the user?

Behavior

- Do you think the behavior of the model's users could change by using the model?
- How was determined how the user interaction with the model is shaped?

Adaptation

- Is the interaction between human and ML model monitored and how?

Dynamic change

- What changes over time could have an impact on the model performance?
- How do you deal with these changes?
- Is there monitoring in place for the ML system and what is monitored?
- How is determined what should be monitored?

Downstream impact

- How was data collected and by whom?
- How is data quality ensured?
- Is the model output or the final decisions used as data for another model?

Accountability

- Can predictions made by the model be reproduced and how?
- Are the final decisions reproducible?
- How is responsibility for model output and decisions arranged?

Directions for a sociotechnical guide

- Do you follow a standardized workflow or process throughout the ML lifecycle or for specific parts?
- Do you document steps taken throughout the ML lifecycle and choices made?
  - if yes, what happens with this documentation?
- How is dealt with responsible ML in the ML lifecycle?

Appendix C

Overview external stakeholder interviews

Besides the interviews with internal stakeholders within the banks that have a role in the two ML use cases, representatives of independent external stakeholders have been interviewed. These are: Waag, Bits of Freedom, Amnesty International, Platform Bescherming Burgerrechten, Privacy-First, The Dutch Data protection Autority and De Nederlandsche Bank.

Below, the interview protocol used for the interviews with representative of these organisations. The questions are grouped in themes. In each interview, every theme is touched upon, but the asked questions differ from interview to interview.

C.1 Interview questions

Introduction

- Can you tell us about your organisation's role in Machine Learning within banks?
- How do you view the use of Machine Learning within banks?
- How does your organisation relate to banks that develop and use Machine Learning models?
- How does your organisation relate to organisations that develop Machine Learning models for banks?

Role in specification

- At which moments do you have a role? For example, before starting to develop a model, during the development or during the use of a model within an organisation?

Sociotechnical dimensions
Misspecification: these are vulnerabilities that arise because the ML system is not properly specified as part of a larger sociotechnical system, consisting of people, processes, and other technical systems

- Do you recognise misspecification as a possible problem?

Machine error: Errors that can occur in the models or the output of models and how possible errors are dealt with

- Do you recognise mistakes and how they are dealt with as a problem?

Interpretability: How interpretability is handled in model development and interaction

- Do you recognise problems due to lack of interpretability?
- Do you recognise problems in the interaction between model and user?

Behavior: The influence the implementation of an ML system can have on people's behavior

- Do you recognise the influence on human behaviour as a source of problems?

Dynamic change: Vulnerabilities that may arise over time due to changes when an ML model is in use

- Do you recognise changes over time?

Downstream impact: how vulnerabilities arising at one point in the model development process can impact later points in the development process or the wider organisation

- Do you recognise downstream impact?

Directions for a sociotechnical guide

- To what extent do you find these categories of vulnerabilities relevant for the development and use of ML systems?
- Do you think I missed any vulnerabilities or categories in my research?

Accountability: How the accountability of decisions taken on the basis of ML systems is organised.

- Do you think accountability can be a problem of ML systems?

Validation of dimensions and vulnerabilities

- To what extent do you find these dimensions relevant for the development and use of ML systems?
- Do you think I have missed any vulnerabilities or dimensions in my research?
- What consequences do you think that not properly addressing vulnerabilities in these categories could have and for whom?
- To what extent do you think banks are aware of these vulnerabilities in the development and use of ML systems?
- To what extent do you think developers of ML systems are aware of these vulnerabilities in the development and use of ML systems?

Directions for a sociotechnical guide

- How should banks using ML expect to deal with these vulnerabilities?
- To what extent do you think banks are currently developing and using ML in a responsible way?
- Do you think that the regulations that are currently in place are sufficient to ensure the responsible use of ML?
- Do you think that the supervision that is currently in place is sufficient?
- Do you have ideas on how to stimulate a responsible and safe use of machine learning, from regulation and supervision or in other ways?
- Do you have ideas about which perspectives should be involved in the development of responsible ML systems?
- Which perspective do you have yourself?
- Do you have ideas about how involving these perspectives could look like in practice?
- Do you already know examples of how this happens in practice?

Appendix D

Insights and recommendations for Deeploy

This appendix provides insights and recommendations for the Deeploy platform and broader activities that the company may take on based on the research outcome. Table 8.2 provides an overview of the guidelines, the implications of them and the accommodation the Deeploy platform provides or may provide in the future.

D.1 Guidelines that are supported by the Deeploy platform

The Deeploy platform supports the implementation of (parts) some developed guidelines, namely guideline 3, 5, 6, 7, and 8. Below, an elaboration on the support Deeploy currently provides for taking up each of these guidelines in ML use cases follows.

Supporting guideline 3: Enable the identification, addressing and mitigation of vulnerabilities in the sociotechnical specification
Guideline 3 prescribes that in the sociotechnical specification, effort should be put in identifying, addressing and mitigating vulnerabilities that can emerge in the sociotechnical ML system. Vulnerabilities can emerge throughout the phases of the sociotechnical ML lifecycle. The Deeploy platform can support both detecting vulnerabilities that emerge in the operations stage of the sociotechnical ML system and support mitigating certain vulnerabilities. For example, the vulnerability function creep, as described in Section 4.5.3, can be mitigated by Deeploy as it is registered which models are deployed in Deeploy. Further, data shifts and concept drift (see Section 4.5.18 and Section 4.5.19) can be detected by monitoring mechanisms in Deeploy such as feature distributions and prediction volumes. Besides, Deeploy mitigates two other vulnerabilities. First, lack of reproducibility of decisions (Section 4.5.22) is mitigated using Deeploy, as the model output and the final decision made by the end-user are made reproducible. Second, Deeploy acknowledges that explainability solutions are no general attributes (see Section 4.5.13), and provides a variety of explainers that can be used simultaneously to meet the requirements of different stakeholders involved, such as the end-user of the ML system, the second line of defence within the organisation or the people about whom a decision is made using the ML system.

Supporting guideline 5: Create feedback channels for different stakeholders during development and operations of the sociotechnical ML system
Guideline 5 prescribes that feedback channels should be created for team members and other stakeholders during development and operations of the sociotechnical ML system. Deeploy partly accommodates this by providing a feedback channel for end-users during operations. In Deeploy, end-users can view and evaluate predictions. An end-user can be provided with the autonomy to either validate or overrule a prediction. In the case that the end-user thinks that the prediction is incorrect, he can overrule the prediction with the 'correct' value. Moreover, he can explain why the prediction should be the 'correct' value. This feedback can be used to further improve the ML system, for example by adding missing features to the model or further optimize the model. Caution with this functionality is recommended, as this is strongly related with potential vulnerabilities. For example, if the end-user is put to a high workload, he might adapt his behaviour to be able to finish his work. This way, it may be attractive to simply validate each prediction. In turn, the data scientists receive no feedback, which seems like the ML system is functioning very

well, while this might not be the case.

Supporting guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical ML system in operation
The sixth guideline prescribes that monitoring and evaluation mechanisms should be in place for the sociotechnical ML system in operation. Deeploy provides monitoring mechanisms for the ML system. Besides the feedback channel that can function as a monitoring mechanism of model predictions, Deeploy offers technical monitoring, concept drift monitoring and event monitoring. Technical monitoring tracks whether the ML system creates predictions, whether errors occur, what the runtime is, and the activity measured over time. Concept drift monitoring tracks concept drift and data shifts. Last, event monitoring logs changes in the metadata of a deployment, such as changes in deployment owner of a model deployment.
Further, alerts can be given to responsible stakeholders based on the monitoring mechanisms. For example, a threshold can be set for the concept drift monitoring. If the concept drift exceeds the threshold, an alert message can be sent to the responsible data scientist, after which the ML system should be re-evaluated.

Supporting guideline 7: Verify and validate the sociotechnical ML system before operationalizing
Guideline 7 prescribes that the sociotechnical ML system should be verified and validated before it is operationalized in the organisation and becomes 'business as usual', Deeploy facilitates deployment and streamlines integration of the ML model in its application environment, which makes it possible to validate the ML system in its empirical context rather easy. As such, Deeploy enables the validation of the ML model's functioning, the operation of the ML model as part of the ML system and application and the interaction with other technical system. Furthermore, the interaction with human-decision makers can be validated by means of the feedback mechanism described above.

Supporting guideline 8: Establish transparency of the sociotechnical ML system
Guideline 8 prescribes that transparency of sociotechnical ML systems should be established. Deeploy contributes to transparency by explainability, reproducibility and traceability. First, Deeploy provides the possibility to use explainers to make ML models and predictions explainable. As explained in the paragraph on guideline 3, different explainers for different stakeholders can be used. Further, to be able to be transparent about which model is used for which prediction, Deeploy provides reproducibility of all predictions. Last, to be able to trace back potential error due to changes made, Deeploy provides traceability of changes made.

D.2 Potential future additions to Deeploy platform

To further improve the support of Deeploy to the guidelines, this section presents potential future additions the company could make to the Deeploy platform. The additions are presented per guideline they support.

Potential addition 1: Implement governance constraints
Deeploy already displays which models are deployed in the organisation. This feature can be expanded by putting governance constraints. These constraints could for example require the approval from appointed stakeholders before a new model, new model version, or the model with updated training data can be deployed. Furthermore, a constraint can be that an ML model that is developed and deployed for one decision-making process, cannot directly be deployed for another decision-making process. This way, an organisation can keep grip on where in the organisation the ML model and predictions are used. This is relevant to prevent downstream impact of a model to work its way through the organisation, without knowing. Also, misspecification is prevented, as portability of an ML system between different sociotechnical contexts is prevented. This addition will thus contribute to mitigating potential vulnerabilities within these dimensions. These governance constraints also contribute to the transparency of the sociotechnical ML system, as it establishes governance for what decision-making process and ML system or its outputs are used within the organisation. As such, this potential addition both contributes to guideline 3, Enable the identification, addressing, and mitigation of vulnerabilities in the sociotechnical specification, and guideline 8: Establish transparency of the sociotechnical ML system.

Potential addition 2: Provide feedback mechanisms for people affected by ML de-

cision
Deeploy already provides a feedback mechanism to provide feedback to ML system predictions, that can be used to improve the ML system and detect potential machine error. This feedback mechanism is currently mainly meant for business experts, for example the human decision-maker that uses the ML system output in their final decision in DSSs. This functionality could be further expanded by creating feedback mechanisms for different stakeholders involved. This expansion requires the possibility to create different feedback roles for different stakeholders, which could be given the possibility to give different types of feedback. Stakeholders that could be provided a feedback mechanism are, besides the business expert, data scientists, compliance officers, risk officers, and people about whom the ML predicts something. For example, people about whom a decision is made, can state whether they agree or disagree with the decision made, and why. Or they can give feedback to the explanation that is provided with the decision. This feedback can be used to detect potential incorrect predictions and revisit a decision made about the person, or can be used for potential improvements in the model and explainability. It is important to consider for every context which stakeholders should be provided with a feedback mechanism, and what action should follow from feedback provided by every stakeholder. Adding this feedback mechanism contributes to accommodating guideline 5, Create feedback channels for different stakeholders throughout the development and operations of the sociotechnical ML system, as it provides additional feedback channels for different stakeholders. Further, it contributes to guideline 8: Establish transparency of the sociotechnical ML system, as it enables different stakeholders to express concerns regarding decisions made.

Potential addition 3: Expand the alerts to responsible stakeholders
Automatic alerts can be sent to responsible stakeholders in Deeploy. This feature can be expanded to serve more purposes. Alerts can already be given to a responsible data scientist in case a monitoring threshold is exceeded, as described in Section D.1. New responsibilities, besides the deployment owner that can be defined already, could be added to other relevant stakeholders in the multidisciplinary team. For example, a compliance officer, a data engineer, could be appointed to a model deployment, along with certain specified responsibilities. For example, alerts can be sent to the multidisciplinary team periodically when the sociotechnical ML system in operation has to be re-evaluated. The team could decide how often this is needed. Further, an alert could be sent to the compliance officer when the data scientist wants to deploy a new version, which requires approval from the compliance officer. This addition contributes to guideline 1: Establish a multidisciplinary team at the beginning of the sociotechnical ML lifecycle, as it accommodates the responsibilities different team members have for the sociotechnical ML system in operation. Further, it contributes to guideline 6: Specify monitoring and evaluation mechanisms for the sociotechnical ML system in operation, as it can alert stakeholders on needed evaluation.

D.3 Recommendations for solution design activities

As Deeploy is a technical tool that accommodates mainly the ML system in operation, several other activities have to be performed to establish and operate a sociotechnical ML system effectively and safely. Deeploy potentially wants to support the activities around the operations in the Deeploy platform. The other guidelines and other directions in the guidelines addressed above mainly require organisational effort. It is recommended to Deeploy to use these guidelines for these solution design activities around the Deeploy platform. A comprehensive overview of the guidelines and directions that can be followed can be found in Chapter 8. A few specific recommendations to the solution design activities are presented below:

- Establish a multidisciplinary team to work on the solution design, consisting of stakeholders within the customer organisation and within Deeploy, to be able to include all needed perspectives to the design. In the beginning, involving the rights stakeholders can be a challenge, and situations may occur in which important stakeholders are missed at the beginning of an ML use case and involved later on. It is key to document these learnings, to be able to improve the establishment of multidisciplinary teams, and the challenges it may bring over time. After more ML use cases, it becomes more and more known for similar types of use cases or industries, which types of stakeholders are important to involve and how stakeholder tensions can be solved.
- Define the system boundaries carefully to include all relevant technical and social subsystems and interactions in the solution design. Keep attention to the impact integrating the

sociotechnical ML system might have in the organisation and beyond.

- As described guideline 3, potential vulnerabilities that may lead to harm should be identified, addressed and mitigated in the sociotechnical specification. It is recommended to use the identified vulnerabilities in Chapter 4 as first inspiration to perform this activity. As vulnerabilities may be overlooked or emerge over time, it is vital to capture these and use these insights to be able to identify them sooner in next use cases.