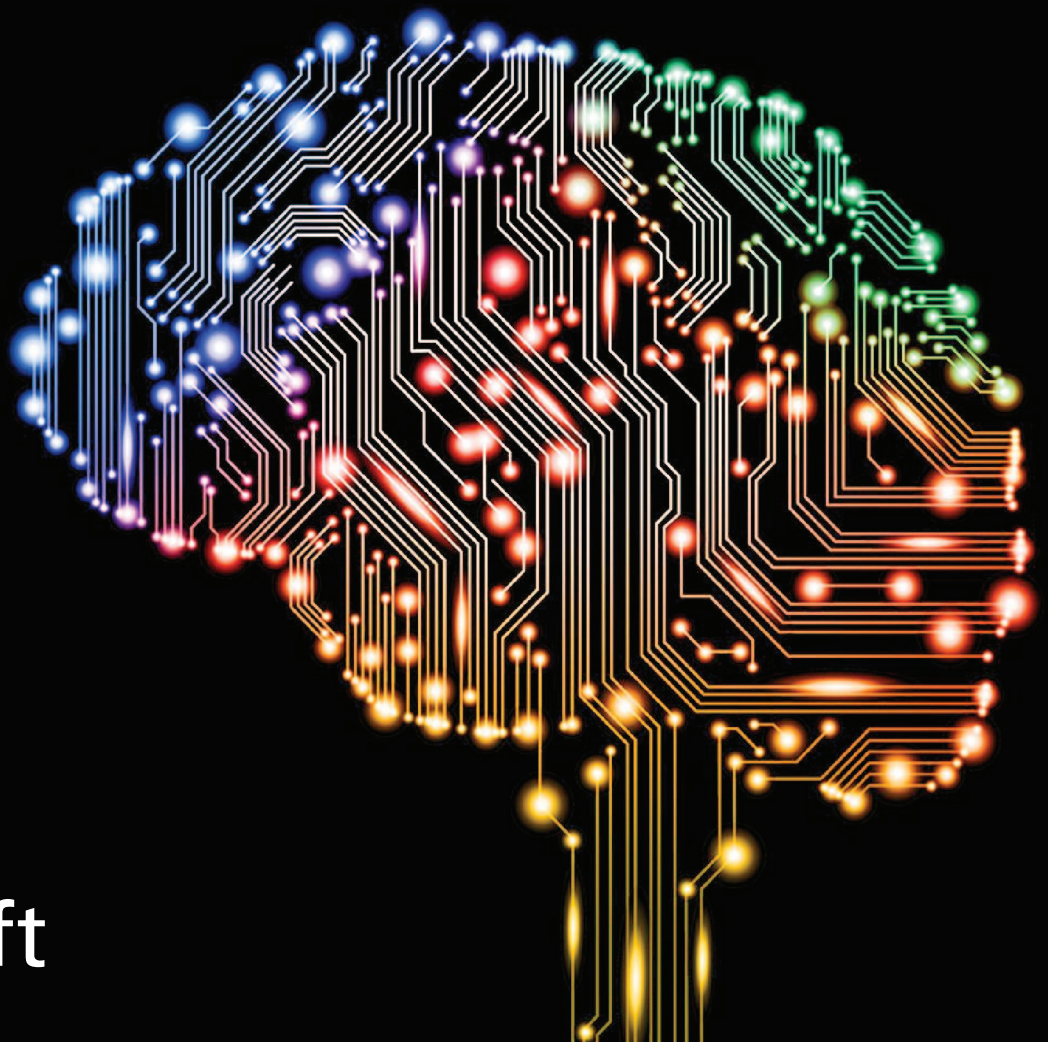
Situating Explainable AI in the socio-technical context

A system safety inspired approach to operationalizing explainability

Alexander Kempen

Delft University of Technology



TU Delft

Situating Explainable AI in the socio-technical context

A system safety inspired approach to operationalizing explainability

by

Alexander Kempen

Student Number: 4394909
Defence: December 20th, 2022

Graduation committee

Chairperson:	Prof.dr.ir. (M.F.W.H.A) Marijn Janssen
First Supervisor:	Dr.ir. (R.I.J.) Roel Dobbe
Second Supervisor:	Dr. (B) Ben Wagner
External Supervisor:	Dr. (J) Jie Yang
Exteral Supervisor:	Dr.drs. (E) Edo van Uitert
Faculty:	Faculty of Technology, Policy, and Management, Delft

TU Delft

Acknowledgements

The master thesis in front of you tries to define explainability within Machine Learning based decision support systems. It does so by situating explainability in the socio-technical context and by establishing a user-centered operationalization of explainability. I hope that this thesis can serve as a starting point for expanding the view on explainability and getting a step closer to capturing all elements that constitute the realization of explainable Machine Learning systems. The research provides a novel way of thinking about explainability, as a system property that must be controlled for.

It must be said that the perspective on explainability prosed within this research has been around me all along. This is during my time at Delft University of Technology (TU Delft), where the explanations of the Professors are controlled for the intended audience through exams. An interesting analogy isn't it? And I must say, during my time at the TU Delft I did not encounter such a tough exam as this thesis. However, I look back at a wonderful period filled with interesting courses and (a lot) of extracurricular activities for better or worse. Finalizing my thesis made me realize that I can proudly say that I grasp the concepts of systems theory and socio-technical systems. I realized even more that this perspective is now more important than ever as systems are becoming increasingly complex. I hope that I can combine both in my future work and will try to encourage everyone along the way to do the same.

The realization of this thesis would not have been possible without my Graduation Committee. Therefore, I would like to thank Marijn for being my Chair and providing me with clear directions and suggestions. Also, I would like to thank Ben for your feedback, support, and interesting discussions. I want to thank Jie for your useful insights and for providing me with feedback on the technical elements of my research. Then I would like to thank Edo, you have consistently provided me with extensive feedback and our weekly meetings brought me a lot of confidence and insights. I enjoyed the time at the bank during the initial phase of the research and getting our hands dirty with some programming (or falafel wraps during lunch). I enjoyed your guidance and finalizing my work at the demonstration and evaluation session. Then, I would sincerely like to thank Roel. To date, we have spent some hours together and I have enjoyed it remarkably. Thank you for our interesting (off-topic) discussions and guidance. You have inspired me and shown me the importance of a systems view within the existing technology-driven society.

This final chapter concludes both my thesis and my time in Delft. It has been a ride and luckily I have gained some knowledge, experience, and a lot of friends. Lastly, I would like to thank my family and friends for their unconditional support. Writing this thesis has been a rewarding experience and I hope you will enjoy reading it.

Alexander Kempen,
December 5th, 2022.

Executive Summary

Artificial Intelligence (AI) is present in our everyday lives and the recent mass adoption is fostered by factors such as the elimination of constraints on computing power, increased availability of data, improved algorithms, and improved and extended open-source libraries that enable knowledge sharing. AI systems have contributed to scientific breakthroughs and are broadly commercialized within areas such as healthcare, finance, and governments, this places us in the third AI 'summer'. AI is becoming the powerhouse behind many decision-making processes and besides the ability to process and find patterns in high-dimensional data, elements such as the ease of scalability and reproducibility, speed of computations, and consistency can make these processes more efficient and fair. However, with the introduction of AI in decision-making processes that revolve around human subjects and data come questions of ethics and morality. These concerns are supported by incidents where AI decision support systems adversely affected society and people. One of the main reasons for these incidents was that the inner workings and decision rules established were not clear and known but still relied upon. With the unconstrained technical complexity of today's systems come the difficulty of understanding and explaining these systems. Now that the dust around AI is settling there are more and more concerns about explainability and critics questioning the over-inflated expectations of AI, potentially bringing us back to an AI 'winter' again.

There are sectors highly in need of the performance and efficiency AI brings as is the case with Financial Crime Detection within banks that try to fight money laundering and terrorist financing by building Transaction Monitoring systems. Criminals today are leveraging the technology-driven society to their advantage, according to the International Monetary Fund (IMF) the figures are staggering and the aggregated size of money laundering is between 2%-5% of the global gross domestic public (GDP). Banks are obliged by law to detect and mitigate these practices and are rapidly setting up anti-money laundering and combatting the financing of terrorism systems that use high-end technologies such as Machine Learning, a sub-field of AI. Banks have not been putting enough effort into establishing these systems and are currently under high pressure as they are receiving fines up to hundreds of millions of euros with even the management being sued. There is a clear need for AI to empower the decision-making process of Transaction Monitoring, however, there is also legislation on the need for explainability as these processes involve human subjects and data. Explainable AI is the field concerned with trying to make AI understandable to humans. While efforts have resulted in significant improvement in research and practical methods of Explainable AI, there is an urgent need for additional research and empirical studies. The academic research gaps identified in this thesis show that Explainable AI is still in its infancy and is mostly approached with a technocentric perspective while not being focused on the audience the explainability is actually intended for. Next to this, there is no structured approach to defining and establishing explainability in dynamic complex systems that involve people, institutional, and organizational elements. Lastly, there are limited empirical studies that investigate the needs, usage, and risk of explainability in complex systems. This research is performed at the Transaction Monitoring department at a large bank and tries to fill these gaps by performing Design Science Research (DSR) with the goal of answering the following main research question:

What does explainability entail in the socio-technical context of Machine Learning based Transaction Monitoring systems?

The research tries to define and address explainability in the socio-technical context within the Machine Learning decision support systems of Transaction Monitoring. It does so by performing an extensive literature review and by conducting semi-structured that try to collect empirical knowledge within local practice. The goal of this research is to expand the definitions and view on explainability by incorporating the social, organizational, and institutional elements that influence explainability. Next, the goal is to develop a method that can help practitioners approach explainability in a structured manner taking the audience into account while applying this socio-technical perspective.

The literature review showed existing limitations and discrepancies of technical Explainable AI methods such as the limited ability to accurately represent the model behavior due to the inability of handling complex feature interactions. Next to this, the literature shows that explainability is highly influenced by latent dimensions such as end-user expertise, time available for interpreting the explanations, and the domain in which the model is deployed. The empirical study showed that the technical limitations are often known by the designer but not by the end-user and that the design and maintenance of explainability approaches are often guided by the intuition of the designer and not by the recipients, this shows a misalignment of mental models of the system. Also, factors that influence explainability practices such as time pressure, keyperson risk, lack of documentation, and control for understanding have been observed as causes for hazards within the local practice. These elements have resulted in flawed design requirements, inadequate decision-making, and even the discarding of entire models. When this is unanticipated this can lead to risks and losses within the organization but can also affect clients (e.g. natural persons). The empirical study showed that explainability approaches have limitations but that inexplainability is mainly caused by processes revolving the design, maintenance, and usage of explainability approaches. These processes must be designed cautiously to avoid the misalignment of mental models, asynchronous evolution, and flawed design requirements. This shows that explainability is highly influenced by socio-technical factors and that it must be controlled for between the interaction of components (or people). The factors that influence explainability approaches show similarities with elements influencing safety in complex systems. Therefore, to control for explainability decades of experience from system safety theory will inspire this research on how to design and control for explainability.

System safety theory uses systems theory and is built upon three main constructs that are safety design constraints, the hierarchical safety control structure, and process models. This research uses inspiration from the first two elements to designing a method that can operationalize and control for explainability while taking the intended audience into account. The method is built upon the established pattern of explainability in the socio-technical context. The method tries to fill the research gap by incorporating the audience and considering socio-technical elements. Next to this, the method will add to the empirical studies on explainability and provide the local practice with an actionable and structural approach to design and control for explainability. The method is pictured in Figure 1 and consists of five steps but should be considered an iterative process.

The user-centered method for operationalizing explainability takes on a socio-technical perspective and can provide requirements for design choices, in addition to this the method shows how these requirements can be satisfied and controlled by instantiating control structures. The method has been demonstrated and evaluated within the bank by providing a workshop using

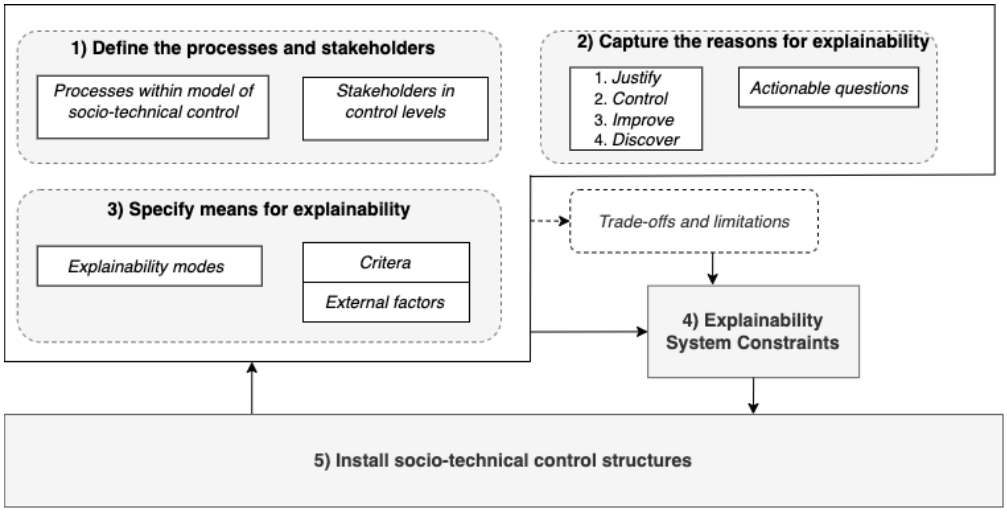


Figure 1.: Method for a user-centered operationalization of explainability

a Toy Case during a focus group. Practitioners experienced the method as useful and actionable, also the method provides broader perspectives and insights on explainability and invites the discussion of dilemmas and questions. The practitioners added that the method could be further refined by focusing on additional guidance on the control structure because the method assumes prerequisite knowledge of systems theory and system safety theory.

Learning from systems theory, however, lower-level behavior can only be influenced by establishing constraints on top of the hierarchical levels. Therefore, this research added five recommendations for organizations that use Machine Learning based decision support systems on how to approach explainability system-wide and from a socio-technical perspective. These recommendations are as follows:

1. Embed explainability in the company culture
2. Create an explainability development plan
3. Operationalize explainability using a user-centered approach
4. Install structured communication channels
5. Avoid complexity and re-think the actual objective

The research contributed both to academics by establishing a novel approach to situate explainability in the socio-technical context and using concepts from system safety theory to establish explainability approaches. Next to this, the actionable method incorporates this view and operationalizes explainability from a user-centered perspective which provides additional empirical study for explainability as well as a structured approach for practitioners. The main limitation of this research is that the feedback on the method has not been incorporated within an iterative design process due to time constraints. Therefore, the main suggestion for future research is to further evaluate and refine the method. A final recommendation for future research is to explore and validate the positioning of explainability as an emergent system property that is in need of socio-technical control using concepts from system safety theory.

Contents

List of Figures

List of Tables

Acronyms

1. Introduction

1.1. The Rise of AI in High Stake Decision-making

In the past decades, there has been a significant increase in the capabilities and applications of Artificial Intelligence (AI) technology (Bryson, 2019). Although AI was already introduced in the 1950s and has been a part of the industrial repertoire since the 1980s, recent improvements in computing storage capabilities, data processing abilities, and the ever-increasing availability of data have contributed towards societal mass adoption in the past decade (Duan et al., 2019). Especially the ability of AI systems to surpass human ability at human pursuits has made global headlines as Deep Blue (Campbell et al., 2002) defeated former World Chess Champion Garry Kasparov in 1997 or the more recent victory of AlphaGo (Silver et al., 2016) against professional Go player Lee Sedol in 2016. One of the earliest and most promising applications is the usage of AI in expert systems that are designed to inform, assist, or automate human decision-making processes (Duan et al., 2019; Edwards et al., 2000). These systems have the capability of increasing the efficiency and accuracy of decision-making processes in many important areas such as health care, public services, or finance which could lead to better diagnoses for patients, improving the productivity of governments, combating terrorism, and many other applications. As AI is being widely adopted in the public sector due to the premise of a more effective, neutral, and low-cost handling of public administration, concerns are being raised as the outputs of these systems are relied upon as decisional aides to human decision-makers (Alon-Barkat and Busuioc, 2022).

The AI systems that have the best performance on such real-life applications are often inherently complex and prone to sub-symbolism that, in contrary to simpler rule-based symbolic systems, do not exhibit clearly defined human-readable relations between input and output (Ilkou and Koutraki, 2020). The inability to understand the internal workings of such systems can exhibit dangerous implications when relied upon for decision-making in critical fields. In such cases, the rationale of a decision often matters, where for example in jurisdiction lawyers use explanation as their primary tradecraft (Lipton, 2018; Selbst and Barocas, 2018). The raised concerns become particularly relevant given the impact of existing failures or malfunctioning of AI systems in highly consequential socio-technical areas. Such malfunctioning happened in 2020 with the Dutch Childcare Benefit Scandal, it started when the Dutch tax agency had decided in 2011 to introduce a sociotechnical system that assessed applicants before paying out benefits. The algorithm used in the system assigned a risk score to applicants and labeled particular cases as potential fraudsters, these cases were then further investigated by officials. However, the officials were heavily epistemically dependent on the system as the only output generated was a probability representing suspicion (or not) without supporting evidence or alternative sources of information (Buijsman and Veluwenkamp, 2022). Such a system is designed to support human experts in making informed decisions by leveraging the capabilities of, in this case, Machine Learning (ML) which is a subset of AI techniques. Combining data-driven and human decision-making can improve the accuracy of a model significantly (Ostheimer et al., 2021). However,

when the involved humans have no insights into the rationale behind the decisions of the system or are not domain experts it can be harder to contest the outcome and their added value is little to none. Eventually, the algorithms from the government were investigated and were by itself even described as discriminatory and filled with institutional bias, consequently, 47.217 parents have appealed for financial compensation (van Huffelen, 2021). The Dutch Childcare Benefit Scandal has shed a light on the importance of understanding models that are deployed within a high stake decision-making process, stipulating the need for a clear view of the internal relations between processed information and outcomes.

1.2. A Socio-Technical View on AI

The meaning of the socio-technical view is shaped by the combination of people, the social interactions they have, the resources they may use, and the technology itself that enables them to act. van de Poel (2020) state that a socio-technical system can be characterized by three main building blocks: technical artifacts, human agents or social interaction, and institutions. This view can and should be applied to AI systems as these systems function on the interplay between technology, humans, social, and technical norms (van de Poel, 2020). When AI systems are adopted, e.g. decision-making processes, they become nested in a larger socio-technical system that can affect human behavior or lives. Therefore, it is important to adopt a socio-technical systems perspective within the development and deployment of AI systems to ensure values such as fairness, transparency, and explainability (Benk et al., 2022). By taking on a socio-technical view on AI both the technical components (i.e. the code and data) and the socio elements (i.e. stakeholders and society in which the system is deployed) will be considered together as a whole, making up the AI system (Dignum, 2019; Sartori and Theodorou, 2022).

1.3. AI in Society and the Challenge of Explainability

The (mis)use of AI can have a tremendous impact on society as discussed earlier with the example of the Dutch Childcare Benefit Scandal. This impact is noticeable in a multitude of layers in society and even called for a change in the institutional discourse on algorithms. The following Dutch cabinet, whereas the third Rutte cabinet resigned due to the implications of the Dutch Childcare Benefit Scandal, included plans to instantiate an algorithmic watchdog in the governmental coalition agreement for 2021/2025 which is intended to safeguard public values by checking for transparency, discrimination, and randomness in algorithms (Rutte et al., 2021). The societal urge for developing a greater understanding of deployed algorithms and models is cross-border and also adopted by the General Data Protection Regulations whereas Article 14(2)(g) stresses that automated decision-making systems should be able to provide meaningful information about the logic involved (European Union, 2016). Last year the European Commission introduced a proposal for a regulatory framework on AI called the AI-act. The pioneering proposal stipulates the need for AI systems to be sufficiently transparent, explainable, and documented. These legislative courses of action show that regulators are acknowledging the importance of developing the right institutional environment in order to safeguard public values and prevent undesired consequences of AI.

While the institutional environment is being set out, there already exists a technical field and study that is concerned with increasing the explainability and interpretability of AI, this is the

field of eXplainable Artificial Intelligence (XAI). As a starting reference point, the definition of the term XAI is given by Gunning et al. (2019) as follows:

"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners".

1.4. The Use of AI in Financial Crime Detection

Recent development in XAI may contribute towards the realization of these regulations and guidelines. XAI systems aim to make existing models more intelligible to humans and foster the development of human-understandable models by providing explanations while maintaining a high level of performance (Gunning et al., 2019). Although more explainable rule-based algorithms are being used in a variety of sectors there is a need for more complex models that exhibit higher performance, this shows the societal shift from merely focusing on technological development to a need for increasing the means to adopt and keep up with to the rapid technological changes (Kile, 2013). A sector that is heavily reliant on rule-based systems is the financial sector, especially within the domain of Anti-Money Laundering (AML). But, fraudulent activity is becoming more difficult to trace in the modern age and increasingly forcing banks and financial institutions to incorporate more complex algorithms that perform more accurately and efficiently. Apart from the fact that banks and financial institutions want to prevent harm from happening to their clients they are obliged by law to act against suspicious behavior and must install transaction monitoring systems and know your customer processes. Next to this, they must still adhere to the more recent regulations and laws described earlier to create transparent, explainable, and documented models and processes which result in a conflicting situation.

1.5. Academic Knowledge Gap and Research Problem

The use of AI systems has enormous potential to increase efficiency and effectiveness for automating and supporting decision-making processes in various fields. However, the use of these systems and specific models encounters many challenges that can impact individuals and organizations (de Bruijn et al., 2021; Sun and Medaglia, 2019). XAI may solve some of these challenges and can aid in safeguarding public values, making XAI increasingly important to all users or those affected by AI systems. Barredo Arrieta et al. (2020) addresses that explainability is one of the main barriers impeding the practical adoption of AI. However, there are still major challenges that XAI needs to face to unlock its potential.

The first main challenge for XAI is that explanations are ambiguous. Research from Barredo Arrieta et al. (2020) concludes that there is not yet a consensus on what exactly explainability entails within the AI realm and stresses the need to define this. A unified concept of explainability will create common ground and must convey the needs that are expressed within the community. Barredo Arrieta et al. (2020) proceed by defining the main goal of an explanation is that it tries to inform a certain audience. Therefore, Barredo Arrieta et al. (2020) argues that the audience should be involved in the definition of explainability. Barredo Arrieta et al. (2020) propose to define explainable AI as:

Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.

Within AI systems there are many stakeholder groups that are in need of different types of explanations. Pedreschi et al. (2019) describes that an explanation should be "meaningful", where the meaningfulness of an explanation depends on the context, complexity, intent, and to whom it is presented de Bruijn et al. (2021).

Another challenge is that current explanation methods and research in XAI are mostly focusing on technicalities, using an algorithmic-centered approach. Angelov et al. (2021) state that although a lot of improvements have been made on the technical side of XAI, which poses certain advantages, there is still a lack of methods that provide clear explanations understandable to humans. de Bruijn et al. (2021) tap into this by stating that XAI is assuming a certain level of expertise which is often not met by the public making the explanations open to interpretation and making the user unable to assess or probe the validity of the explanation. This results in the fact that there is still a lack of understanding on how to define and incorporate explainability within a process that involves a multitude of stakeholders expanding the view beyond merely techno-centric.

Lastly, most of the time explanations and explanation methods are used in complex systems, such as organizations, which result in the fact that there are additional environmental factors to take into account. Explanations should be validated and the accuracy of explanation methods must be controlled. To determine whether an explanation is successful, it is necessary to measure whether something has been understood (Barredo Arrieta et al., 2020). Validation is necessary to assess whether an understanding is passed on and if the explanation is still matching the actual process it is trying to explain. Validation and evaluation are still one of the main open challenges within the field of XAI (Barredo Arrieta et al., 2020). Explanation, as described earlier, must involve meaningfulness and this should be validated by the intended audience.

Society calls for an increased understanding of complex ML models in order to trust and deploy these in high-stake decision-making processes such as in the field of financial crime detection. The collaboration, apparent in such decision-making processes, between human and machine intelligence can achieve a synergy as the human intellect can be augmented and achieve goals that are unreachable by one or the other separately (Akata et al., 2020). Sarkara (2022) describe that explainability may be *the* wicked problem of AI, as there are multiple possible approaches to the problem explanation and each of them may be unique. Within the usage of AI for decision-making processes, multiple stakeholders are involved each having their own desires and requirements that may change over time. Wicked problems are ill-structured, change over time, and have many factors and conditions all embedded in a dynamic social context with multiple actors (Rittel and Webber, 1974). As there are no normative criteria for the solution of a wicked problem, understanding the wicked problem is the actual problem, because wicked problems have no stopping rule (Rittel and Webber, 1974). From a scientific perspective, there are still important challenges to face for XAI. This results in both a societal and scientific need for a deeper understanding of the wicked problem of explainability in a dynamic complex system that uses AI as a support for decision-making processes. This results in the following **research problem:**

Despite the recent technical developments of XAI, there is still a lack of knowledge on how to approach XAI from a socio-technical user-centered perspective where explanations are meaningful and can be validated

This thesis aims to solve the research problem by gathering empirical data on explainability within high-stake decision-making processes of financial crime detection using a socio-technical

lens. Insights will try to define explainability from a socio-technical perspective as well as define reasons and requirements for different stakeholders within the process of financial crime detection. In order to gather this data interviews will be conducted with a variety of stakeholders within the decision-making process of Transaction Monitoring (TM), a subfield of financial crime detection. The gained insights will try to fulfill the research problem and will build a foundation for the Design Science Research (DSR), which aims to define and operationalize a user-centered approach to explainability using a socio-technical perspective. This research will be performed in collaboration with a large Dutch bank within the area of financial crime detection, specifically TM, and will have a duration of approximately six months.

1.6. Research Objective and Main Research Question

The research proposed in this paper will dive into this problem by researching possibilities to increase explainability within the decision-making process of TM, by doing so it will try to fulfill both the social and scientific demand for further development of the interaction between human and machine intelligence. The research objective is to operationalize a user-centered approach for approaching explainability from a socio-technical point of view. The research will hereby respond to the scientific challenges XAI is facing and answer to the societal need for research on explainability within decision-making processes. This leads to the following overall **main research question** of this master thesis:

What does explainability entail in the socio-technical context of Machine Learning based Transaction Monitoring systems?

.

This research will contribute towards explainability practices that take the audience into account and social, technological, and institutional elements. This will provide research on explainability that tries to include a socio-technical perspective rather than a technocentric perspective. Next to this, the research will position explainability as an emergent system property that must be controlled in complex dynamic systems. By investigating both a great body of literature and performing an extensive empirical study this research will contribute to academics and society.

1.7. Relevance to the CoSEM master's Program

The research will be a final work to complete the Complex Systems Engineering and Management (CoSEM) Master's program at the University of Technology Delft. The CoSEM Master's program is characterized by having a multidisciplinary approach with the objective of designing interventions in socio-technical systems by doing research on complex issues in real-world decision-making processes. This research situates explainability within the socio-technical context by covering the technical abilities and limitations of explainability approaches and by investigating social and institutional elements that influence explainability needs and practices. The research is performed at a bank and tries to alleviate a real-world problem. The research applies actor analysis and will investigate both institutional and organizational elements of Transaction Monitoring systems and explainability. The research is in line with the CoSEM guidelines as it will try to develop an intervention (method) that is situated in the dynamic complex system of Transaction Monitoring. This method is built upon constructs from system safety theory and

systems theory, whereas the latter is on of the main underpinnings of the CoSEM master. By trying to develop this method, a systems engineering approach will be applied as well as process management strategies. The research will cover values originating from both the public and private domain taking the view of multiple stakeholders into account.

1.8. Thesis Outline

This research will start with elaborating on the research method and approach used, outlined in Chapter 2. Then, Chapter 3 will describe the application environment and map the process and development of TM as well as the existing legislation and techniques applied to battle money laundering. Then the knowledge base will be established on explainability and XAI approaches together with the current challenges and limitations in Chapter 4. The knowledge base will be further developed by researching systems theory and system safety theory in Chapter 5. The empirical study will be performed and elaborated in Chapter 6, discussing the explainability approaches used in TM together with the stakeholder reasons for explainability, risks, and limitations. Next, Chapter 7 will present the method for a user-centered operationalization of explainability taking on a socio-technical view. Chapter 8 will reflect on the insights gathered and provide recommendations on how to approach explainability within organizations. Finally, Chapter 9 will present the main findings of the research together with the developed artifacts and provides the limitations of the research as well as recommendations for future research. Lastly, Chapter 9 will include a personal reflection on the research process as well as a piece of short advice to the CoSEM dean.

2. Research Approach and Methods

2.1. Design Science Research

The research takes on a DSR approach in order to answer the main research question. Design science is defined by Johannesson and Perjons (2014) as *"the scientific study and creation of artifacts as they are developed and used by people with the goal of solving practical problems of general interest."*. Within DSR practical problems are addressed by the creation of an artifact, which represents an object made by humans that supports people when practical problems are encountered in practice. DSR tries to produce knowledge that is built upon an existing scientific body of knowledge and uses empirical data from local practice dealing with the practical problem. Figure 2.1 displays the types of knowledge received and produced when using a DSR approach from Johannesson and Perjons (2014). As can be seen in Figure 2.1, the knowledge produced by DSR should have a scientific contribution to the research community and general practice contributions for local practices. The outcome of a DSR is both the artifact and the contextual knowledge about the artifact.

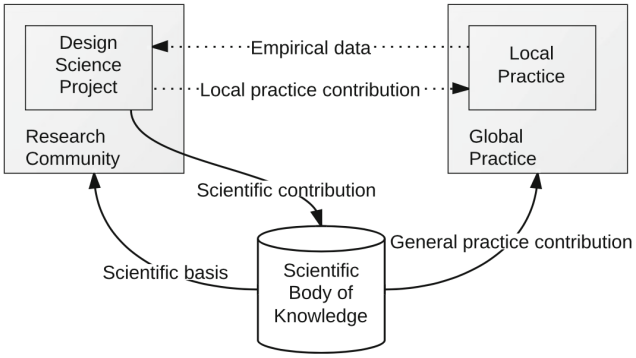


Figure 2.1.: Contributions of design science research described by Johannesson and Perjons (2014)

This thesis will follow the knowledge process depicted in Figure 2.1 by starting with gaining empirical data gained from both experiences and semi-structured interviews with local practitioners working with ML models at a bank within TM. The second main source of information will be from an extensive systematic literature review on XAI tools used for ML to serve as the scientific basis for the DSR. The scientific base and the empirical data will provide the information necessary to design the artifact.

2.1.1. Designing an Artifact

Artifacts can be straightforward and focus on a specific practical problem. However, often artifacts are embedded in systems that involve other artifacts, humans, and the relations and norms that govern their interactions. As described in Section 1.2, an AI model deployed within a decision-making process in society can be viewed as a socio-technical system due to the fact that technical models, stakeholders, institutions, and norms are involved. Within this system, the different roles, desires, and needs of stakeholders contribute to the complexity of the system. This results in a need for an artifact that is able to guide and coordinate the actions of the involved actors. A full description of artifacts and the different types are provided in Appendix D

2.1.2. Design Science Method Framework

In order to perform the DSR a structured approach will add value to the process. The approach will be based upon the three-cycle view of DSR from Hevner (2007). These cycles represent actions in DSR to achieve the contributions. The first focus will be to describe the environment which consists of an application domain, earlier described as a local practice. Therefore, the first cycle is the relevance cycle describing the application domain with its existing actors, institutions, and technical systems. Next to this, the problems and opportunities within the domain will be described that can lead to requirements that serve as input to the design. The next cycle is the rigor cycle which aims to provide past knowledge from experience and expertise in the application domain and from existing artifacts and processes found in the application domain. In addition to this, knowledge will be produced from scientific theories and methods. The input of the design cycle is the requirements from the relevance cycle and the design and evaluation theories and methods from the rigor cycle. Within the design cycle, the artifact is constructed and evaluated. For the evaluation, focus groups will be used to evaluate the artifact. For this thesis, the application domain of financial crime detection focusing on the use of TM models within banks will be researched. Figure 2.2 shows the research approach applied to the three-cycle view from Hevner (2007). The research questions are allocated within the defined cycles to structure the research process.

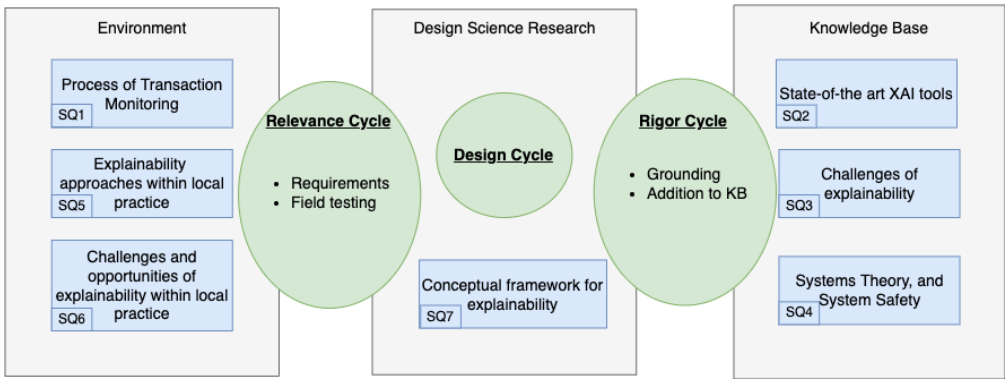


Figure 2.2.: Research approach conveyed in the design science research three cycle view from Hevner (2007)

2.2. Research Strategy and Sub-questions

The research strategy will offer high-level guidance on the activities involved to conduct the DSR. The research strategy will consist of three subsequent phases and is based upon the adopted DSR three-cycle view pictured in Figure 2.2. Within each research phase, sub-questions are posed that are subsequently structured to try to solve the main research question and contribute towards the design of the artifact. The research phases together with their sub-questions will now be elaborated down below and are pictured in Table 2.1.

Label	Phase	Sub-question
SQ1	*Phase 1*	What drives the need for well-developed Transaction Monitoring systems?
SQ2	*Phase 2*	What are the important elements necessary to define good explainability?
SQ3	*Phase 2*	What are the limitations and challenges of current explainability practices?
SQ4	*Phase 2*	What concepts and strategies of system safety theory can aid explainability practices?
SQ5	*Phase 3*	How are banks currently addressing the needs for explainability within the development and operations of TM systems?
SQ6	*Phase 3*	What challenges arise regarding explainability within the process of TM taking on a socio-technical view?
SQ7	*Phase 4*	What method can guide the operationalization of explainability within ML based decision support systems, taking on a socio-technical view?

Table 2.1.: Sub-question per research phase

2.2.1. Research Phase 1: Exploring Transaction Monitoring using a Socio-technical Approach

Phase 1 is within the relevance cycle and will be about setting out the environment of the local practice of the decision-making process of TM within banks. This phase will first explore the TM system by defining the decision-making process, development, institutional environment, involved actors, and technology used in TM. In particular, it will focus on TM models used to combat AML. First, the TM system will be defined by mapping the system using a socio-technical perspective, this will be done in order to gain empirical knowledge on the actors involved, their actions, behavior, and information needs and requirements. These aspects within the system of TM will be defined by answering the first sub-question:

SQ1. What drives the need for well-developed Transaction Monitoring systems?

2.2.2. Research Phase 2: Discovering Explainable AI together with its Challenges and an Exploration of Systems Safety Theory

Phase 2 will consist of the rigor cycle and set out the knowledge base to build a theoretical grounding on the existing explainability approaches. Explainability approaches can be applied to various ML models, only the explainability approaches applicable to TM will be considered. The most significant XAI approaches will be discussed and reflected upon. This will result in an overview of both technical and non-technical approaches that try to increase explainability. This knowledge will be used as a base to understand and further investigate the possibilities for approaching explainability. Sub-question two will investigate explainability practices and will try to do so by answering the following question:

SQ2. What are the important elements necessary to define good explainability?

The upcoming section will research the limiting factors of the explainability approaches and the general challenges found in the literature. This will give insight into determining what is still missing or needed to fulfill the gap between the desired state and the current state. This results in the following sub-question:

SQ3. What are the limitations and challenges of current explainability practices?

The upcoming section will research system safety theory, a theory that applies systems theory and control theory to ensure safety in complex socio-technical systems. This sub-question will investigate whether there are concepts or existing strategies that can potentially be adopted to ensure explainability. This results in the following sub-question:

SQ4. What concepts and strategies of system safety theory can aid explainability practices?

2.2.3. Research Phase 3: Identification of the motivations, approaches, and challenges for explainability in local practice, taking on a socio-technical view

Within the local practice, there is a multitude of entities involved in the process of operating, designing, and regulating TM systems. The involved actors may have different needs and requirements which even can be conflicting, therefore, it is important to map these out and extract knowledge of the involved actors on their needs and reasons for explainability. In addition to this, it is important to investigate the current efforts made by the local practice on approaching explainability. The existing explainability approaches will be discussed that are used within the TM system itself as well as within the development of TM systems. This will result in an answering sub-question five:

SQ5. How are banks currently addressing the needs for explainability within the development and operations of TM systems?

Looking at the needs of the stakeholders and to what extent the bank tries to serve and fulfill those needs, gaps and conflicting interests within explainability requirements of TM systems can be determined. These challenges and gaps should be determined in order to improve explainability within the system, this will be done by answering sub-question seven:

SQ6. What challenges arise regarding explainability within the process of TM taking on a socio-technical view?

2.2.4. Research Phase 4: Designing a Socio-technical Method for Approaching Explainability

The last phase will consist of the design cycle. Now that the knowledge base is defined together with the environment of the local practice, the artifact can be designed. This will be done by using the knowledge base and the requirements from the environment to develop a method that takes on a socio-technical approach to explainability using concepts from system safety theory. The challenges for explainability within TM are defined together with the needs and desires of the involved stakeholders. Next to this, the current state-of-the-art explainability approaches relevant for TM have been investigated. In addition to this, concepts and strategies from system safety theory show how to design control structures to constrain emerging system properties. What is left is to combine this knowledge and develop a method for a user-centered operationalization of explainability within general ML based decision support systems. This method will be demonstrated and evaluated by using a Toy Case within a focus group at the bank. This will result in answering the final sub-question:

SQ7. What method can guide the operationalization of explainability within ML based decision support systems, taking on a socio-technical view

2.3. Research Methods

Research methods provide guidance on a more detailed level and complement the research strategy (Johannesson and Perjons, 2014). The research methods will describe how to collect the necessary data for the research strategy and sub-questions. The methods will be described according to the different phases of the research strategy.

2.3.1. Phase 1: Desk research

For sub-question one a lot of insights are needed on the practical processes occurring in AML and TM within banks. To achieve this, document analysis and desk research will be performed on the internal documents at the banks that reveal embedded knowledge on the aspects of the TM system and the development process.

2.3.2. Phase 2: Systematic Literature Review and Grey Literature

To build a solid knowledge base a systematic literature review will be conducted that will provide data for answering sub-question two and three. The literature review will, among other added values, give theoretical insights on what is known and what may be missing in the field of XAI focusing on approaches applicable in TM (Wee and Banister, 2016). A comprehensive overview of the existing literature can lead to identifying existing challenges, gaps, and possibilities from literature and will produce definitions for key concepts used in this research. The

process of the systematic literature review has been described in Appendix A and the eventual body of literature is presented in Appendix B. Grey literature, in the form of commercial books, will be used to develop a deeper understanding of technical ML concepts and specific XAI methods. Next to this, the literature will be seeking to gain a deeper understanding of systems theory, control theory, and system safety theory in order to answer sub-question four.

2.3.3. Phase 3: Semi-structured Interviews and Data Analysis

Next to this, for answering sub-question five and six interviews will be conducted. The interviews will follow a semi-structured format, this will allow for using an open set of questions where the order may vary. This structure is beneficial for investigating complex issues, as the respondents are not bound to a specific protocol making them unrestricted in answering the questions (Johannesson and Perjons, 2014). Also, a semi-structured format stimulates interaction with the respondent, by mixing open and closed questions to extract answers that are as informative as possible. The interviews will be documented through audio recording, and field notes, and will be transcribed after they have occurred. After transcription, the interviews were coded using ATLAS.ti. Coding allows for a qualitative data analysis that can help in answering the sub-questions. The codes of the interviews can be found in Appendix O. The interviews will try to gain information on the explainability needs of the stakeholders in TM, focusing on sub-question five. Next to this, the main focus will be to extract information that will try to answer sub-question six. Within these questions, information will be retrieved on the current approaches for explainability and explainability challenges at stake within TM. Lastly, in order to gain a deeper technical understanding of TM and the existing XAI tools used within the local practice, focusing respectively on sub-question five, data analysis will be performed on the output of the TM models. The data analysis will be performed in Python and uses Azure Databricks to experiment with existing practices of technical XAI methods on the model output.

2.3.4. Phase 4: Combining retrieved insights

The last phase will combine the retrieved insights from the previous phases in order to answer sub-question seven and does not necessarily have a specific method allocated. Within this phase, a design process will be used to develop an artifact (method) that aims to pose a solution to the main research question.

2.4. Research Flow Diagram

A Research Flow Diagram is created that captures the relation between the research strategy, sub-questions, research methods, and the accompanied deliverable each phase should produce and can be seen in Figure 2.3.

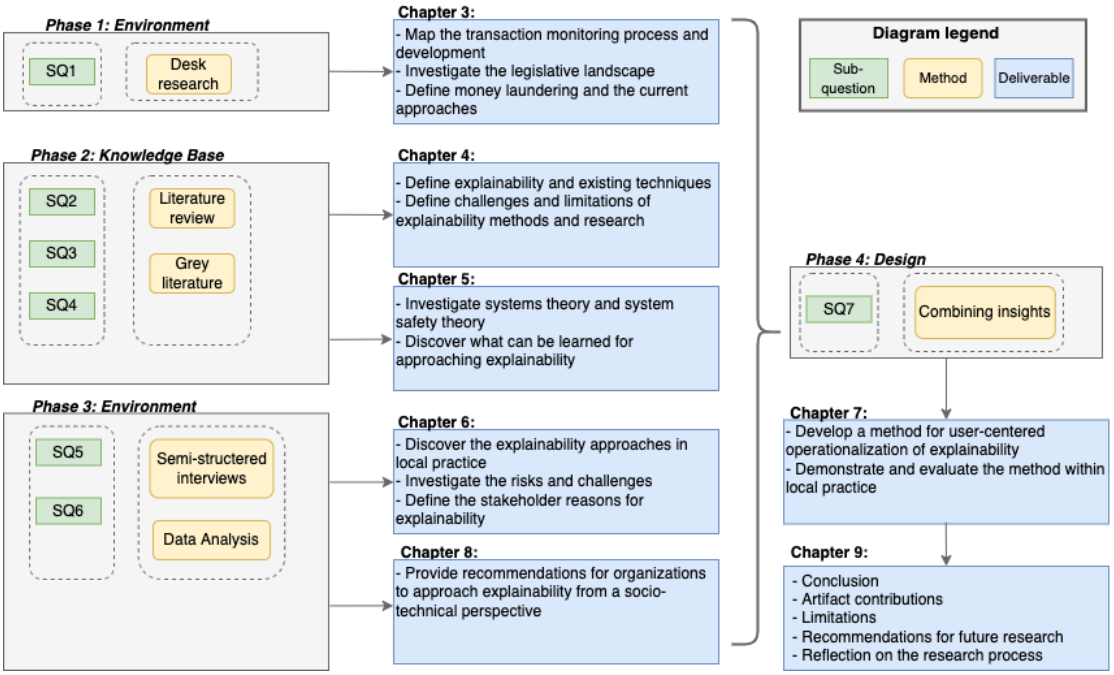


Figure 2.3.: Research flow diagram

3. Detecting Financial Crime

This Chapter will explore the application environment and is situated in the first phase of this research. The Chapter will map the TM landscape by investigating the decision-making process of TM, the development, existing legislation, and the current ML techniques used in TM. This Chapter will close off with Section 3.6 that tries to answer the first sub-question, which is as follows:

SQ1: *What drives the need for well-developed Transaction Monitoring systems?*

3.1. Financial Crime Detection and Anti-money Laundering

Financial crime has become more accessible with the rise of the internet and computers, Hollywood styled bank-robbing or foraging can now be performed by anyone with an internet connection at home and has become mostly non-dependent on geographical location. Although financial crime has been around since the establishment of currencies (or even before), the technical advancements becoming available to humans with malicious intent resulted in the fact that the tools and behavior of criminal activity are becoming more difficult to detect and combat (Nicholls et al., 2021). This can be observed as the World Economic Forum stated that financial crime is a multi-trillion dollar industry of which less than 1% is ever frozen or seized by regulatory agencies (World Economic Forum, 2022). Next to this, the coronavirus pandemic (COVID-19) has impacted the global financial system heavily resulting in a tremendous increase in the use of digital financial services which brings an additional set of new challenges in the war against financial crime (Zhu et al., 2021). This shows that now more than ever there is a need for mitigating and detecting financial crime. Due to the growing threat against the established financial systems and the economy Europol, the European Union Agency for Law Enforcement Cooperation has prioritized the fight against financial crime for the past decade (Europol, 2022a). The definition of financial, or economic crime according to Europol (2022b) is as follows:

"Economic crime, also known as financial crime, refers to illegal acts committed by an individual or a group of individuals to obtain a financial or professional advantage. The principal motive in such crimes is economic gain.".

There are multiple types of financial crime such as money laundering, corruption, widespread counterfeiting, fraud, and tax fraud schemes that target individuals, countries, and companies (Europol, 2022c). Classifying types of crime can be hard as it mostly entails a combination, for example with corruption there is often the case of money laundering involved for example. Within this research, the focus will be on money laundering, because most crimes that have financial gain as a goal involve money laundering. Money laundering is described by the European Directive (EU) 2015/849 as the conversion or transferring of property that is knowingly derived from criminal activity for the purpose of concealing or disguising its origin with the goal of evading legal consequences (European Parliament and the Council, 2015a).

Within the Netherlands financial institutions are obliged, by the Wet ter voorkoming van witwassen en financieren van terrorisme (Wwft), to act against money laundering and terrorist financing and install appropriate AML and Countering the Financing of Terrorism (CFT) instruments. Money laundering and terrorist financing are often categorized together because they both exhibit similar transactional features and patterns. Money laundering is most often concerned with concealing the origin of funds to disguise their criminal nature. With terrorist financing, the funds may be legitimate or proceedings from criminal activity, or both, but rather focused on concealing the destination and eventual use of the funds. However, it is also important for terrorists to conceal the source of the funds so that it remains available for future transfers and financing activities (Schott, 2006). The focus of this research will be on AML which is categorized into two main activities, namely the Know Your Customer (KYC) process and TM process. The focus of this research will be on TM, with TM money laundering patterns can be detected on a transaction level by monitoring transactions and investigating potential suspicious client behavior that is in line with these patterns. The concept of money laundering, as defined above, is relatively simple but the techniques used to disguise the original nature of the proceedings are becoming more advanced and the money laundering process may involve different payment service providers, using multiple intermediary payment service providers, and using different financial instruments coming from or going to different countries.

The general process of money laundering and financing terrorism can be described by three main activities: placement, layering, and integration. First, the resources must be acquired, these can either be legitimate assets or cash from criminal acts. Regarding the latter, these ill-begotten proceedings come from predicate offenses that profited the criminals. Such predicate offenses can be the act of selling illegal substances, bribery, or theft, and are motivated by the desire for profit in any form (Sharman and Chaikin, 2009). An overview and description of the processes involved in money laundering and financing of terrorism have been provided in Appendix E. Money laundering and terrorism financing occur everywhere in the world but may be less constrained in countries with complex financial systems. Or in countries with lacking or corrupt governmental regimes and ineffective or even non-existent mitigation measures and infrastructure (Schott, 2006). Another critical aspect that makes battling money laundering or the financing of terrorism increasingly harder, is the displacement and use of financial institutions hosted in multiple countries throughout the processes involved. In such cases, criminals are benefiting from the challenges, mostly legal, in cross-border information sharing between different jurisdictions.

3.2. Legislative Obligations on Money Laundering and Terrorism Financing

Keeping up with the fast pace environment of innovating technologies is crucial for legislation to stay ahead and prevent unlawful behavior (Silva, 2019). Collaborative action has resulted in the instantiation of the Financial Action Task Force (FATF), which is an intergovernmental organization initiated by the G-7 countries to develop policies and set standards to combat money laundering and encourage AML and CFT practices. The FATF has developed a list of recommendations that encapsulate the technical and legal definitions of money laundering set out in the *Viennna Convention* and the *Palermo Convention* (Nations, 2000; United Nations, 1969). The forty recommendations of the FATF set the standard for fighting both money laundering and terrorist

financing and constitute a framework for detection, prevention, and suppression. These recommendations, or standards, from the FATF are implemented within participating unions and countries. The European Union has incorporated these standards in Directive (EU) 2018/843, which is the 5th AML Directive aiming to battle money laundering along with other financial crimes. The EU legislative measures also include Regulation (EU) 2015/847 which sets out rules on AML that are legally binding for every Member State (European Parliament and the Council, 2015b). Regulation (EU) 2015/847 Article 1 states that *"Money laundering, terrorist financing, and organized crime remain significant problems which should be addressed at Union level"*. Each Member State has to act on the Regulation and implement its own national laws to rule on the goals described in the 5th AML Directive. Within the Netherlands, the European standards are implemented in the Wwft and the Sanctie Wet (SW) which is set up by the Ministry of Finance and the Ministry of Justice and Security. The Wwft is the Dutch Anti-Money Laundering and Anti-Terrorist Financing act, enforcing payment service providers to prevent, mitigate, and act on financial crimes such as money laundering and terrorism financing. The Wwft prescribes two main categories for AML and CFT which are TM and KYC.

3.2.1. Know Your Customer and Transaction Monitoring

De Nederlandsche Bank (DNB) and the Autoriteit Financiele Markten (AFM) are the responsible regulatory instances in the Netherlands supervising the correct implementation of the Wwft by all institutions that fall under the law's reach. Within the Wwft there are two core obligations prescribed for financial institutions. The financial institutions that fall under the ruling of the Wwft are investment firms, investment institutions, undertakings for the collective investment in transferable securities, and financial service providers. Within this research, the focus will be on the latter and specifically banks. The first main obligation is installing a thorough client investigation procedure, also known as KYC. The KYC process can be seen as customer due diligence and is established due to the fact that banks are obligated to track the flow of money. To identify the origin of funds, they require and record information on the payer's identity, the goal of the transfer of funds, and the intentions of the client. The second main obligation is to report unusual transactions to the Financial Intelligence Unit - the Nederlands (FIU). In order to do so, banks should install monitoring systems for all the transactions they process. The FIU is an independent government body and is an active member of international collectives such as the FATF. Appendix F Figure F.1 shows an overview of the most important institutional entities and how they are related in regard to the active legislation, standards, and guidelines.

3.2.2. Risk-based Approach

Within the recommendations of the FATF, setting the global standard for AML and CFT, approaches on how to effectively tackle money laundering and terrorist financing are described for financial institutions such as banks. Within the updated recommendations in 2012 the FATF included that in order to strengthen global safeguards and protect the integrity of the financial system, a Risk-based Approach (RBA) provides an essential foundation of a country's AML and CFT framework (FATF, 2014). Within their recommendation, the FATF writes '*The application of a RBA is not optional, but a prerequisite for effective implementation of the FATF standard*' (FATF, 2014). In order to develop and implement the RBA, existing risks together with the potential mitigating factors should be identified.

Within the Netherlands, the common method used by financial institutions to define and assess risks is through applying a Systematic Integrity-risk Analysis (SIRA). The recommendations for an RBA approach are adopted nationally and are included in the Wwft, which describes that institutions must categorize clients within certain risk levels based on the nature and impact of their associated risks. Financial institutions define these risks themselves and must take appropriate mitigating measures. The risk categories differ from low- to high-risk en must be based on objective and recognizable factors (De Nederlandsche Bank, 2020). The DNB states that the most important aspect of risk categorization is that the decisions and considerations are made in a systematic and consequent manner which allows third parties and regulators to understand and assess the process. In addition to this, DNB writes that by implementing and periodically revising the SIRA institutions can recognize, accept (based on their risk appetite), or avoid existing vulnerabilities. The SIRA analyzes the integrity risks within financial institutions and consists of four steps: risk identification, risk analysis, risk control, and risk monitoring. The process is cyclic and continuous whereas new findings in risk monitoring can lead to (re)defining new integrity risks for example. The SIRA method is based on a holistic risk-based manner that prioritizes cases exhibiting risks with high impact, allowing for more intrusive procedures. Such an approach ensures efficient allocation of resources due to a better alignment between the risk detected and the measures taken to prevent or mitigate money laundering practices (Silva, 2019). An overview of the SIRA process is displayed in Figure 3.1.

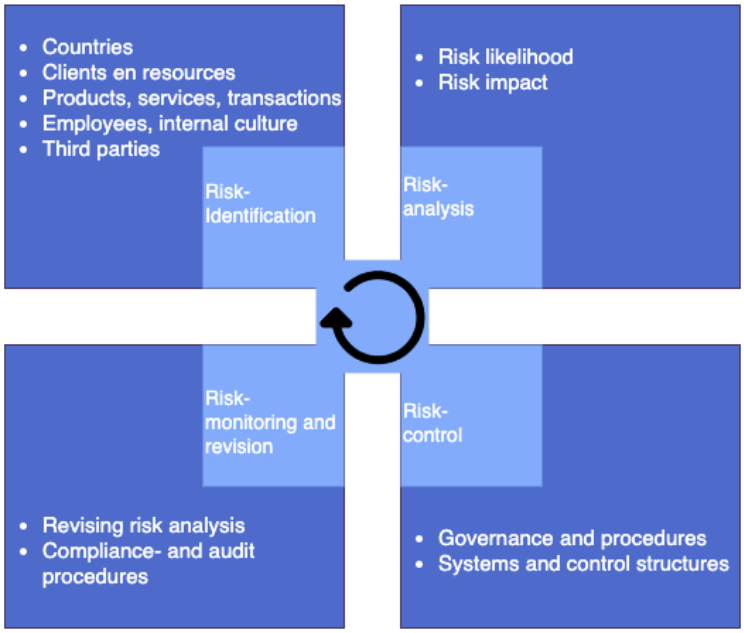


Figure 3.1.: Overview of the systematic integrity risk-analysis interpreted from De Nederlandsche Bank (2020)

The foundation of the SIRA is set within the initial step, the risk identification. Within the risk identification, the financial institution should determine how, and to what extent, it is vulnerable to integrity risks, such as money laundering and terrorism financing. The defined risks should capture a broader spectrum and should include an institution's clients but also the work-

ing culture of the company, its norms, and procedures. Within this research, the focus, however, will be on integrity risks. When the risk is identified an analysis will determine the impact and likelihood of the risks so that it can be categorized, this step will set the prioritization in the SIRA. The risk analysis is performed by domain experts who can assess the impact of a certain risk, within financial banks these are the business lines. According to guidelines on the Wwft from the DNB, the factors that play a role in identifying and assessing integrity risks are based on the type of clients, services, products, transactions, communication, and countries involved (De Nederlandsche Bank, 2020). The intended nature of the business relationship can be defined through customer due diligence, also referenced as the KYC process. When defining the risk category the combination of these factors is taken into account, each factor consists of its own aspects, for example, defining a client will consist of its operating industry, assets, source of income, and many other characteristics. After the risks are categorized and the potential impact is defined, the bank should complement this with additional insights from relevant internal and external stakeholders, and include typologies set up by international organizations or the national government (FATF, 2014).

Organizations should determine their own risk appetite based on impact and prioritization. They must define which risks are they willing to take and which ones are unacceptable. An example of the need for new mitigating strategies will be when risks are categorized as high-impact while the risk appetite is low. Now that the risks are identified, assessed, and prioritized, the financial institution must put in the effort to control and mitigate the risks by developing and implementing policies and procedures. Procedures include TM to check for suspicious patterns in transactions or find inconsistencies in the expected behavior of the customer or client (FATF, 2014). The final step is to monitor the risks and assess whether current procedures and policies are effective and whether new trends or international standards can change the impact of existing risks or give rise to new ones. The financial institution should be resilient to such changes and revise the SIRA continuously, often when new criminal patterns are found they are included in an event library that can complement the following SIRA cycle. Using the SIRA, financial institutions set up their AML and CFT frameworks using the prescribed RBA. The RBA is focused on specifying which mitigating strategies should be applied in order to mitigate the selected critical risks realizing effective efforts in the fight against crime.

3.3. Benefits of strong Anti-Money Laundering and Countering the Financing of Terrorism Systems

There are many benefits from having a well-established AML and CFT infrastructure covering both societal and financial incentives. Societal incentives are often translated into legislative measures safeguarding national or cross-national interests by preventing and decreasing criminal activity which can impact the economy or its citizens. Next to this, effective AML and CFT systems can also strengthen and improve the businesses of financial institutions.

3.3.1. Fighting Crime and Corruption

The main reason for and benefit of a well-established AML and CFT system is to fight and mitigate criminal activity by making predicate offenses less profitable and denying criminals or terrorists working capital (Sharman and Chaikin, 2009). Increasing the means for combatting

money laundering and terrorist financing will disincentivize criminals to use financial institutions for their activities. An important goal is to confiscate and forfeit the proceedings which will eliminate the potential profits for criminals. In addition to this, it will provide governments with another avenue to find and prosecute criminals who are both producing illegally gained proceedings or are active in the money laundering itself. Strong CFT systems will decrease the financial means for terrorists and will, hopefully, result in fewer acts of terrorism. The main driver to establish this will be to develop a broad institutional framework that includes many predicate offenses which must be acted upon. An example of this is when bribery is included as a predicate offense, this will oblige financial institutions to act against corruption and decreases the ability of criminals to bribe officials (Schott, 2006). Setting up a strong AML institutional framework together with AML systems can effectively reduce corruption, this can be beneficial, especially for developing countries as corruption is one of the greatest threats to economic development and good governance (Sharman and Chaikin, 2009).

3.3.2. Building Trust in the Economic Sector

Positive public opinion plays an essential role in key economic activities, especially in financial activities as it will increase the confidence of investors and consumers (Istrefi and Piloiu, 2020). When financial institutions assure they are putting in efforts to mitigate risks that can result in potential financial losses, consumer trust can be increased. Besides this, mitigating the actual financial losses and risks within the operations of the financial institutions will strengthen the financial stability of the institution. An example of this is with regard to sound KYC practices for risk management, by performing good KYC practices financial institutions can define their exposure and lending risks more accurately when issuing a loan which can decrease potential losses (Schott, 2006). The financial loss of banks is not the most important factor, more importantly, is developing trust from clients that the bank is making an effort so that criminals do not benefit from the economic infrastructure

3.3.3. Fostering Economic Development

Financial crime is one of the main drivers that can adversely affect a nation's business activity (Schlossberger, 2015). For the case of money laundering, the inability to install secure AML systems will allow criminals to successfully integrate, or often invest, their illegal proceedings in the regular economy, properties, or other assets. Criminals seek investments that are easily transferable and do not lose value, however, these investments often do not generate additional productivity for the broader economy and are called sterile investments (Schott, 2006). Most often, these are mostly high-value luxury goods such as art and jewelry, however, research from Schott (2006) shows that criminals even transform productive enterprises into sterile investments for the main purpose of laundering criminal proceedings. Previously, these enterprises generated profit and now become unresponsive to consumer demand resulting in unproductive use for capital. When this is applied on a large scale it can harm the productivity of a country's economy as enterprises with profit-generating purposes and other resources are turned into sterile investments (Schott, 2006).

3.4. Transaction Monitoring

Financial institutions are obliged to take measures to detect, mitigate, and report integrity risks such as money laundering and terrorism financing. As discussed previously, one important aspect is monitoring transactions and transaction behavior from clients. The SIRA defines the unusual and suspicious behavior that can lead to integrity risks and is the foundation of the governance, business processes, and procedures regarding TM. Financial institutions are obliged to report transactions that are linked with integrity risks to the FIU who will further investigate the case. Whenever the institution does not adhere to the duty of alerting the institution and its management can even undergo criminal charges. Financial institutions use software-based TM systems that analyze transaction data in order to generate alerts of suspicious behavior. TM can be based solely on a transaction level, where single transactions can lead to an alert due to its characteristics.

3.4.1. Monitoring Different Types of Transactions

Known Unusual Transactions
As discussed previously, the SIRA internalizes national regulation within local practices, such as banks, and incorporates this in their own governance by defining their set of risks from which patterns and rules can be extracted to define unusual transactions. Unusual transactions detected by the mitigating measures from the SIRA can be referred to as known unknowns, this is because they exhibit risk of which people are aware of (Luft and Ingham, 1961). Within the Wwft concepts of unusual transactions or suspicious behavior remain undefined. This is because these terms are inevitably prone to ambiguity as they are ever-changing and subjective. However, as a reference, the Wwft defines a few common patterns that may be classified as unusual transactions or behavior. An example of this can be when a single high-value cash (e.g. 100.000 euros) transaction is deposited, this is because criminals are known to use cash due to its anonymity and will therefore be classified as a known unusual transaction. Another way to approach known unusual transactions is to define certain transaction profiles for clients based on the expected transaction behavior and usage. Clients can be alerted once they deviate from what is considered within the normal (low-risk) range of behavior. Such patterns, eventually defining unusual behavior, are determined by the banks themselves using objective indicators that can be known.

Unknown Unusual Transactions
Because economic trends and technological innovations are changing constantly, banks and regulators are faced with the challenges of adapting to this ever-changing landscape. Criminals are using today's technology-driven society to exploit all instruments available at their disposal in order to innovate and foster new illegal activities. Therefore, banks and regulators need to be able to detect newly arising patterns in order to battle money laundering. However, situations might exist where patterns and risks can be considered which are unknown and cannot easily be detected. Such patterns and risks cause the definition of unknown unknowns, which are situations that exhibit unknown risks of which people are unaware of (Luft and Ingham, 1961). Such unknown unusual transactions are not included in the SIRA. The unknown unusual transaction can only be detected by constantly innovating the techniques, or setting up advanced ML solutions that are able to detect new uncommon patterns of risky behavior.

3.4.2. Process of Transaction Monitoring

The goal of TM systems is to detect any type of suspicious behavior, this can either be due to the violation of certain thresholds within the business rules or when triggered by more advanced ML models generating an alert. When such an alert is generated an internal investigation will start. A human, or TM analyst, evaluates the alert and can escalate this by filing an internal Suspicious Activity Report (iSAR). Within this step, the TM analyst assesses the alert and documents his considerations and conclusions before he either closes or reports the alert. When an iSAR is filed, an additional internal investigation is conducted after which it is decided whether or not it will be turned into a Suspicious Activity Report (SAR). SAR filings will be reported and handed over to the FIU.

This shows the nature of TM as a decision-making support process, where a human assesses the output of the model and improves the next iteration of the model when it is retrained on recent data. The human in this process is referred to as the TM analyst and ensures additional safeguarding before proceeding to act on the output of the system (Kuiper et al., 2021). The global process of TM consists of three steps that can be divided into monitoring (alerts), investigating, and reporting. Taking inspiration from Nesvijevskaia et al. (2021), who applied the Fraud Management Lifecycle Theory from Wilhelm (2004) on retail banking, an overview has been created for the TM process together with the different steps and actions that can be taken. The high-level overview of the TM process within banks is pictured in Figure 3.2, for an elaborate description of the subsequent steps in the TM process Appendix G can be consulted.

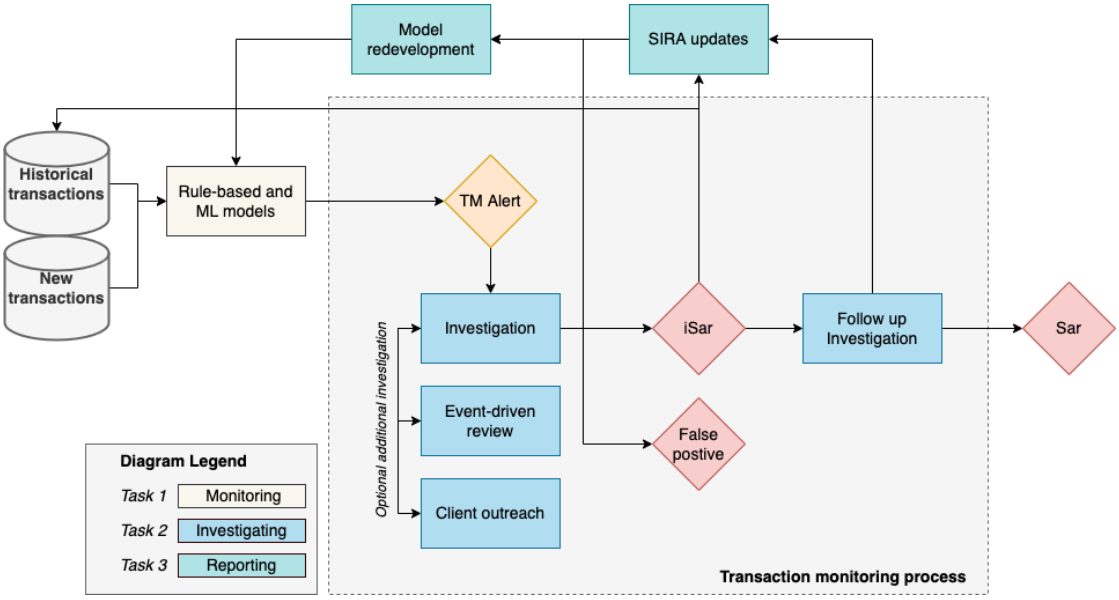


Figure 3.2.: Process of transaction monitoring

3.4.3. Development of Transaction Monitoring Systems

The development of ML models for TM models within the local practice can be divided into four phases. These phases are the initiation phase, design phase, development phase, and produc-

tion phase. Each phase is elaborately described in Appendix H. The development process has an iterative nature and there is a lot of communication between entities from different phases. However, some elements in the process are extremely strict because the models operate in a high-stake domain that can have a severe impact on clients. For example, the testing phase within the development is often labor intensive and can be seen as repetitive, however, the model will never reach production before this step is finished. The development process is pictured in Figure 3.3 accompanied by the main actors involved.

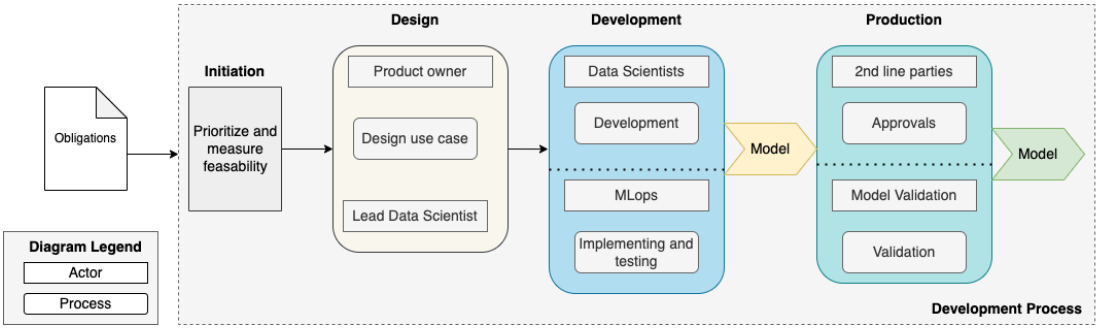


Figure 3.3.: Process of developing transaction monitoring models

3.5. Machine Learning for Transaction Monitoring

For the use case of AML both rule-based monitoring representing symbolic if-then rules and more advanced techniques such as supervised and unsupervised machine learning is used. These ML techniques will be elaborately discussed in Chapter 4. Although, in the past TM was only done by using such rule-based scenarios, nowadays, more advanced ML models are being applied in conjunction with the rule-based methodology (Kuiper et al., 2021). Figure 3.4 shows the interplay of different types of models and how they provide feedback for one another. Figure 3.4 shows that both rule-based and ML models are built upon the SIRA and alert transactions that can lead to iSAR filings. Next to this, it can be seen that more advanced ML models are sometimes used on the output of rule-based models to determine whether the reasoning is correct. Each model works towards the same overarching classification task of detecting unusual transactions and behavior, however, their inner workings and objectives may differ. Within AML there are no requirements on what type of AI system are used, as long as the decision-making and rationale can be explained both to internal stakeholders and to the supervisory authorities. Kuiper et al. (2021) state that AML is one of the main use cases where supervisory authorities allow room for the use of state-of-the-art AI technology and where such systems can be most beneficial regarding the possibility to improve results.

According to Gao and Xu (2009), the two most preferred AML methods for TM are AML topologies and anomaly detection. AML topologies, as described by Gao and Xu (2009), refer to the ability of algorithms to find and report fraudulent cases based on similar historical occurrences. The AML topology methods require historical transaction data whereas transactions are labeled, such label represents whether the transaction is previously defined as a suspicious transaction or belongs to the set of regular transactions. The second well-established AML method for TM is anomaly detection and refers to the ability of an algorithm to detect outliers within the set

of transactions and identify unusual or suspicious behavior deviating from the norm. The following subsections will first discuss the rule-based systems that apply business rules to detect AML typologies. Next, supervised learning will be discussed which uses data to detect AML typologies, and afterward unsupervised learning is discussed which is often used for anomaly detection, also the advantages and existing challenges will be discussed for both these techniques.

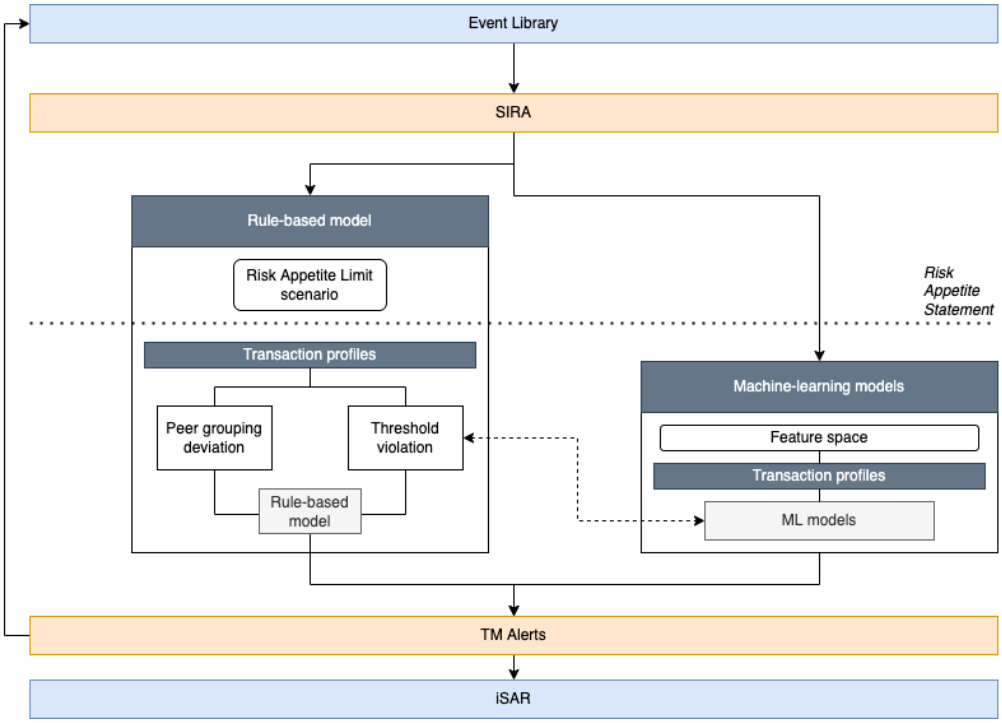


Figure 3.4.: Transaction monitoring model interaction

3.5.1. Rule-based systems

Banks have set out certain types of decision rules to mitigate risks from the SIRA to define suspicious behavior as the term can be subjectively interpreted. Such embedded rules are set out by domain experts implementing their working experience which can be applied within an operating system that tries to detect known unusual transactions (Chen et al., 2018). The system that applies these decision rules, based on thresholds, is called a rule-based system and incorporates expert knowledge by applying pre-generated conditional if-then rules (Hayes-Roth, 1985). Fixed rules are an efficient way to act against extreme behavior that violates the decision boundary, or threshold, and due to their transparency a good method to enforce regulatory guidelines (Gao and Xu, 2009). Rule-based systems can also be applied for AML topology, without the training phase, whenever a scenario or transaction is determined by domain experts and historically proven to be suspicious.

However, there are some pitfalls and challenges that come with rule-based systems applied in AML. These static rules need to be updated or added to adapt to changing trends and data as

more exceptions can arise. While the set of rules is increasing and existing and newer rules need to be integrated, the ability to adapt and perform on dynamic data can decrease (Gao and Xu, 2009). Apart from this, it is extremely difficult to capture all exceptions within the set of rules and to evaluate the performance of the entire set (Gao and Xu, 2009). Lastly, as the volume of transactions keeps growing chances are that different types of data will be used, being semi-structured or unstructured, which hampers the performance of rule-based systems. Next to this, because the rule-based systems are based on domain knowledge it is hard to capture complex patterns. In addition to this, rule-based systems can not handle a variety of rules, this is because they are static and can result in a client staying unalerted whenever he operates just below the thresholds. Eventually, rule-based systems often lead to a high number of false positives.

3.5.2. Supervised Machine Learning

Supervised ML can deal with the growing set of transactions, complex patterns, and dynamic data by learning the decision rules from the data itself without consolidating domain experts. Supervised learning is trained on labeled historical data, whereas the label is often represented by a binary variable differentiating suspicious behavior from non-suspicious. It does so by labeling non-suspicious transactions as examples, hence creating an understanding of regular behavior or transactions. This implies that supervised models only recognize patterns from the past (i.e. the data it has learned from) and have difficulties recognizing changing patterns. However, retraining the model on recent data forms a natural way to make a supervised model familiar with new emerging patterns.

Training a supervised ML model on historical data gives the model the ability to generalize, this enables the model to classify new scenarios with unseen data into the predefined label categories without explicitly being trained on such an event before (Gao and Xu, 2009). Often, the model outputs a score representing the likelihood of suspicious behavior. Whenever such a score violates a certain threshold, set out by the financial institution, an alert will be raised to further investigate the case. Supervised learning is often used for AML topology and often standard open access models for classification (e.g. logistic regression, decision trees) are implemented (Hastie et al., 2009).

3.5.3. Unsupervised Machine Learning

Whereas supervised learning uses labeled data, unsupervised learning can extract patterns from the data without requiring labels. An example is an unsupervised clustering method that can address anomaly detection and has the ability to detect outliers within or between clusters of transactions (Nesvijevskaia et al., 2021). This specific unsupervised learning method classifies data examples into clusters without label information (Gao and Xu, 2009). Often with unsupervised learning, there is no access to labeled historical data, however, when there is the labels are omitted during training and only used when evaluating the cluster formation (Gao and Xu, 2009). The objects within the formed clusters represent similar data instances and are dissimilar to the other clusters. Clustering discovers natural grouping in the data which can potentially reveal hidden patterns represented in unknown unusual transactions for example. When using a clustering algorithm for unsupervised learning the goal is to discover inherent groupings within the data based on certain patterns, for a financial bank this could be the grouping of clients based on their transaction behavior. For applying an anomaly detection algorithm

the goal is to discover rules that describe large portions of the data, within the example of a financial bank this could be that if customers transfer their money to destination X they often do to Y. So, anomaly detection is often used to determine very unusual instances of the data that deviate from these established patterns. Within AML this might indicate that the behavior of a client is very different from those of others and can be an indicator of suspicious behavior. Unsupervised learning benefits from using unlabelled data, refraining from the need for domain expertise to label data appropriately which can be costly and time-consuming.

3.6. Main Findings Chapter 3

The goal of this Chapter was to familiarize with key concepts in the practical environment such as money laundering, financing of terrorism, and the legislative field trying to enforce mitigating approaches such as TM and KYC. Also, the process of TM and the development of TM systems have been mapped. Next to this, this Chapter aims to answer the first research sub-question:

SQ1: *What drives the need for well-developed Transaction Monitoring systems?*

TM can benefit society, financial institutions, and clients but the main societal need for establishing TM systems is to fight crime and corruption by making predicate offenses less profitable and eliminating potential profits for criminals. Next to this, well-developed TM systems built trust in the economic sector which plays a vital role in macro-economic prosperity because trust serves to connect actors within the economy and influence the way they work together and propel growth. Finally, well-developed TM systems can foster economic development and avoid unproductive use for capital. These needs are all vital and recognized by intergovernmental organizations such as the G7 and multiple legislative instances have been initiated. It is widely recognized that AML and CFT efforts, especially due to increased connectivity, are cross-border and require cooperation. Within the Netherlands, the most important rulings on AML and CFT are the Wwft and SW enforced by the DNB and AFM. The Wwft prescribes that banks must implement TM and KYC practices by taking on an RBA. Banks try to do so by developing and describing SIRA risk which they try to mitigate, these mitigating measures consist of rule-based models and ML models. Both types of models work together trying to alert suspicious behavior on unusual transactions. Banks are trying to constantly improve the TM systems by trying to incorporate more advanced techniques that can find both known and unknown unusual transactions.

4. Explainable Machine Learning Systems

This Chapter will try to strengthen the knowledge base and is situated in the second phase of this research. The Chapter will try to answer the second and third sub-question, which are as follows:

SQ2: *What are the important elements necessary to define good explainability?*

SQ3: *What are the limitations and challenges of explainability practices applicable to TM?*

4.1. Machine Learning

AI is the overarching definition for the theory and development of computer systems that try to mimic human intelligence. Specific application of AI are therefore often around tasks that normally require human intelligence such as visual perception, decision-making, and speech recognition. Within the realm of AI there are many different kinds of techniques and systems that try to achieve such tasks. This includes handcrafted knowledge systems which incorporate human-programmed rules which come from human subjects and are automated by machines. Examples of such systems are expert systems which are often referred to as rule-based systems. However, ML also falls within the definition of AI and includes techniques that incorporate machine-programmed rules whereas the knowledge is learned by applying an algorithm to data. ML, consisting of a set of methods, is used to draw inferences from patterns in data and is often applied to make predictions.

Within ML there are many different types of methods which are often classified into three categories based on their type of training: *supervised learning*, *unsupervised learning*, and *reinforcement learning*. Algorithmic improvement is stimulated by multiple factors but mainly due to the elimination of constraints on computing power, the increased availability of data, improved algorithms, and improved open source libraries (Allen, 2020). These trends have powered both the development of new algorithms and improving other methods immensely which has fueled the general re-interest in AI.

Machine learning techniques often surpass human ability in certain tasks, especially in processing and finding patterns in high-dimensional data. This is no problem for ML models but for humans, such data is incomprehensible let alone impossible to find patterns and relations. However, there are also situations in which ML algorithms do not necessarily outperform humans, but there are still great advantages to them such as the ease of scaling, the speed, and reproducibility (Molnar, 2020). When a ML model is implemented it can infinitely be transferred to other machines with low costs and can complete tasks much more easily than humans. Also, ML models are consistent in their results where humans often can vary their decisions based on external factors that do not have any real correlation with the decision at hand. But, the downside of ML models can be that the best performing models can use millions of parameters making it extremely hard, or even impossible, to understand the computations entirely. This

complexity makes it hard to retrieve insights on the data and on tasks the ML is solving to find an answer. The winning models in competitions are often very complex, mostly deep neural networks, or can be a blend of several different models, called ensembles, and are almost impossible to interpret (Molnar, 2020). Within this thesis, the focus will be on ML systems that use supervised, unsupervised, and semi-supervised learning and do not include techniques such as neural networks or deep learning. For this research, the focus will be on ML systems designed to support human decision-making in critical fields using tabular data. Tabular data is defined by any classical dataset, where the data is stored in rows and columns, in which every record shares the same set of features with the features either being numerical, categorical, or boolean values (Guidotti et al., 2018).

4.1.1. Supervised Machine Learning

Supervised learning is currently the most used form of ML and is programmed to learn a function that maps the input to an output based on a set of data that includes the solutions on which the model is trained. The algorithm will be trained to understand the association between the features of a data instance and the desired value that needs to be predicted, this makes the predicted variable the dependent variable and the features of the target the independent variable. X is the notation representing the feature set, which is an independent random variable, and y is the notation for the observation which is desired to predict or forecast. In order to learn the relationship between X and y while being able to predict a new y given an X, which is not in the training set, the model should be trained on data that contains the exact answer to the problem. This implicates a constraint on the data necessary to train supervised models as both the features and true outcomes should be known.

The training phase, where all the values of X and y are known, is used to learn the relationship between X and y and where it can refine its learning and establish the learning rules. The training data set, which contains a representative sample of data points for which X and y are known, is used to learn the relationship between the features and the target outcome. Training an algorithm on the training data set will allow the model to learn a function that maps the features to the target outcome, during training it does so by optimizing toward a predefined objective. Once training is complete, the performance of the model can be evaluated on the test set. The test set uses new unseen data and feeds this into the model while withholding the target variable. This will allow evaluation of the ability of the model to map the input to a correct output by comparing the output of prediction to the actual target variable. In order to predict the target variable for the new unseen features the model uses the mapping function established during the training phase. The most basic form of the mapping function can be given as follows, where Y is the predicted output determined by the mapping function that assigns a class to the input value x.

$$Y = f(x)$$

Supervised learning can be split into two classes depending on the task it is used for, these classes are called *classification* and *regression*. For classification, the supervised model will be assigning the output to a certain class and in the case of regression, the model outputs a numerical value. The performance of a model is optimized during training towards a certain goal, often this can be represented by a function called the objective or score function which tries to minimize the difference between the predicted output and the target variable. Application for supervised learning in TM have been described within Subsection 3.5.2.

4.1.2. Unsupervised Machine Learning

Unsupervised learning is a form of ML which uses unclassified or unlabeled data on an algorithm and allows the algorithm to act on that information without guidance, hence unsupervised. An unsupervised model is used to group unsorted information according to patterns, similarities, and differences in the data. In unsupervised learning, the training procedure is concerned with finding such similarities between the samples from the training data and putting similar items together in sets. During training, a certain threshold is found to find similarity within the raw data and in cases where an instance shares no similarity with other instances, it will become its own set. Although there is no ground truth, the test set can be used to give an unbiased estimate of the performance of the model-building method. During testing the similarity for all instances within a set is calculated after which a test instance is compared to see whether it is similar to at least one item in a set. When the test instance is similar it will be allocated to that set. Applications for unsupervised learning in TM have been described within Subsection 3.5.3.

4.2. Interpretability of Machine Learning Models

Within research there has been not been consensus on the term explainability and interpretability, these terms are often used interchangeably. However, some researchers define interpretability as being able to discern the mechanics without knowing the exact cause (how) and where explainability takes this a step further and refers to the untangling of the reason (why) (Gunning and Aha, 2019; Padovan et al., 2022; Lughofer et al., 2017). To illustrate this with an example, interpretability refers to knowing how water gets to a point of boiling, and explainability is focusing on why water is boiling at a certain temperature. Knowing this ambiguity, this research will choose to use explainability and interpretability interchangeably as do multiple other researchers.

To set consensus within this research the notion of interpretability within the context of ML is used from Doshi-Velez and Kim (2017) which defines interpretability, or explainability, as *"the ability to explain or to present in understandable terms to a human"*, hence, this reflects the degree to which a human can understand the cause of a decision and can consistently predict the result of a model. Doshi-Velez and Kim (2017) argues that interpretability is a step-stone required to meet other important ML desiderata such as fairness, privacy, reliability, robustness, causality, usability, and trust. Because interpretability is defined as the ability to explain a certain phenomenon, presenting a rationale, the term explanation should be defined. Explanations are a means to achieve the ability of interpretation, however, a formal definition of an explanation still remains elusive and debated in different fields of research. But an explanation will be defined as an answer to a why-question posed by an audience (Miller, 2019). The motivation for trying to answer such why-questions seems clear within the ML community based on a different set of motivations. Doshi-Velez and Kim (2017) states that the need for interpretability, and thus explanations, is coming in cases where it is not only important to get the prediction, but also to know how it came to the prediction (the why) because sometimes a correct prediction only partly solves the original problem (Molnar, 2020). Interpretability and explanations can solve such incompleteness as it can show the effects of the gap in the problem formalization. Multiple perspectives, described by Molnar (2020) and Kaya (2022), on the motivation for knowing an answer to the why-question are listed and discussed down below.

- **Human-AI cooperation and acceptance**: explanations can lead to increased understanding of the model and can optimize inference with humans. Also, explanations can lead to the ability to appropriately trust or accept the model.
- **Regulatory compliance and High-risk applications**: explanations can measure whether a model satisfies certain legal requirements, such as the General Data Protection Regulation (GDPR). The GDPR obliges that models do not discriminate on the basis of certain demographics. Such an additional constraint is a part of the model's problem formulation but may not be incorporated into the mathematical objective function of the model. Interpretability can aid in debugging the model to bridge and check for such incompleteness. Explanations can show whether a model is consistent in predicting a certain outcome when instances have similar properties to validate the usage of ML in high-stake decision-making areas. By using explainability it can show that a model is non-discriminatory and not influenced by environmental factors.
- **Safety measures**: ensuring that safety constraints are followed can only be determined when models can be probed to check whether the abstraction the system has learned is error-free and matches the intended behavior. Often for complex tasks, the entire system is not completely testable (Doshi-Velez and Kim, 2017). Explanations can help in finding edge cases that might cause ambiguity for the models' programmed safety constraints which can cause a system failure. An example of this is with the case of autonomous vehicles where interpretability can show that the model defines cyclists as two-wheel objects, however, when a cyclist in real life uses a side bag that covers the wheels this can lead to an error.
- **Model debugging and auditing**: explanations can be used by model developers or auditors to understand why the model makes certain mistakes in order to fix them. Interpreting can help in finding the cause of the error within the model.
- **Scientific understanding**: scientific findings remain unexplored when a model only gives predictions without explanations. When explanations are incorporated this can lead to knowledge discovery of the detection of new relations and interactions within research and potentially find causal factors. Interpretability can extract the knowledge captured by the model and therefore gain new knowledge, the ultimate goal of science.

4.2.1. Reasons for Explainability

The motivations for answering a why-question on the mechanics or usage of ML models is clear and can be generalized. The reasons for explainability can be approached and categorized in four different domains, these reasons coming forth out of the research from Adadi and Berrada (2018) are depicted in Figure 4.1. These four reasons are: explain to justify, explain to control, explain to improve, and explain to discover. Although these domains might share some similarities they are distinct enough to separate the rationale and most importantly can be used to categorize the stakeholder needs for explainability.

Explain to Justify
The first domain stems from the need to justify a decision focusing on reasons why a particular outcome is presented rather than a description of the inner workings of a model (Adadi and Berrada, 2018). Often this perspective is taken when auditing a model checking whether decisions are made ethically, just, and fair. Recent legislation is installed to make sure decisions

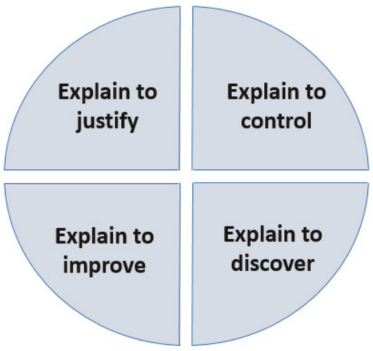


Figure 4.1.: Reasons for explainability, retrieved from Adadi and Berrada (2018)

are not made erroneously preventing biased and discriminatory results. Such need is realized within the GDPR as it includes a "right to explanation" for automated decision-making systems. Explanations to justify are especially an important aspect within high-risk applications and regulated industries that need to be compliant with defined regulations.

Explain to Control
Apart from justifying decisions, there is a need to enable control by creating an increased understanding of the system's behavior. This can aid in testing safety constraints and gaining insights on unknown vulnerabilities and flaws (Adadi and Berrada, 2018). The scope of control can vary from correcting errors on an operational level (debugging) or to a model perspective with regard to safety measures. In these situations, explainability can aid in enabling enhanced control of the model behavior and outcome.

Explain to Improve
A thorough understanding of a model results in the ability to make better improvements because the creators, or users, of the system, know why certain outcomes or behavior are apparent. This can result in improved efficiency and smarter models.

Explain to Discover
Knowledge discovery can only be realized when explanations are installed that can provide information showing new facts or relations. This can only be achieved by explainable models which can aid in increased scientific understanding or in finding undetected patterns and behaviors in domain-specific fields such as AML.

It is important to think about why and when explanations are necessary as they require considerable resources both in the development and human interaction of ML models (Adadi and Berrada, 2018). Adadi and Berrada (2018) states that incorporating explainability might even result in less efficient ML systems, forced design choices, and a bias towards explainable, but less capable and versatile outcomes. On the other end of the explainability axis, research from Doshi-Velez and Kim (2017) and Molnar (2020) state that indeed explanations are not necessary but only when either (1) the model has no significant impact and there are no unacceptable consequences when the predictions are wrong or unexplainable or (2) the problem is well-studied with enough practical experience and validation so that system's decision can be trusted without the need of additional insights.

4.2.2. Scope of Interpretability

The answer provided to a why-question is strongly related to the level at which the question is posed, therefore, in order to define and assess the interpretability requirements the scope should be determined (Miller, 2019). The level on which the explanations are based, or the scope of interpretability, can differ a lot per domain, industry, and type of model. For example in high-stake domain areas, it is important that single predictions or group predictions can be explained in order to analyze the consistency and reliability of the model. The levels of interpretability are pictured in Figure 4.2 and will be described shortly.

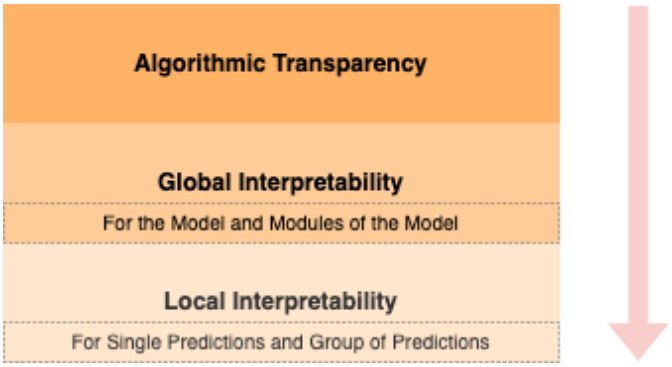


Figure 4.2.: Levels of interpretability

Algorithmic Transparency
The highest level, algorithmic transparency, is concerned with how the algorithm creates the model. This can show how the algorithm learns patterns from the data and knowing what kind of relationships it can learn (Molnar, 2020). Algorithmic transparency only focuses on the knowledge and computations of the algorithm and not on the data used or the eventual model, this can be seen as a preliminary before starting to build a model with an algorithm. Although it is important to know how an algorithm works, it does not concern model interpretability and will not be considered within this research.

Global Interpretability
Moving one level down, the global model interpretability is perhaps the most difficult to achieve in practice. In order to explain the global model output, there should be an understanding on how the model makes decisions based on its features and the learned components such as weights and other parameters (Molnar, 2020). Global model interpretability can help to understand the distribution of the target outcome based on the features which can aid in determining what features are important or which interactions between features. However, it is almost always impossible to comprehend the entire model at once because the human mind is not able to imagine multi-dimensional hyperplanes which are necessary to comprehend the feature space. So, often modules of a model are considered in order to comprehend a model, with linear models these are often the weights that are given the most attention as they are easily interpreted at a parameter level. However, when interpreting on a modular level there is a certain assumption to be made, such as the fact that in the case of linear models interpretation of a single weight can only be done when all other input features remain at the same value. Often this assumption does not hold in real-life applications because weights do only make sense in the context of other features of the mode. However, this still provides some degree of interpretability which can give additional insights.

Local Interpretability
The lowest level is local interpretability which examines single instances and tries to explain why the model predicts the instance for a certain input. Local explanations are often more accurate than global explanations and by zooming in on the individual predictions the global complex behavior can be broken down most of the time into more simple linear or monotonous dependencies on some features.

Single predictions can also be grouped and explained on a global modular level or on a local level. The global modular method will take the group of individual predictions and treat it as if it were the complete dataset using the global methods on this subset. Local individual methods can also be used on every single instance of the group after which they can be aggregated for the entire group.

4.2.3. Latent Dimensions of Interpretability

The dimensions that characterize the importance of interpretability are mostly dependent on the task for which the ML model is designed, as discussed in Section 4.2. This can differ from safety-critical domains to domains that have little to no impact. When considering the interpretability of a model, the set of dimensions should be considered that are dependent on the task at hand can change the view and the importance of interpretability. Building forth upon the research from Guidotti et al. (2018) and Doshi-Velez and Kim (2017) the following task-related latent dimensions can be formulated:

- **Domain:** the domain in which the eventual model is applied heavily affects the type and need for explanations. May this be in a high-stake decision-making area such as healthcare or autonomous vehicles one might want to have an exhaustive list of scenarios and determine whether safety criteria are being fulfilled or not. Also, certain domains or usage of specific data falls under the ruling of certain regulations, such as the GDPR or AML directive. This can also impact the need for explainability.
- **Severity of Incompleteness:** determining the kind of interpretability depends on the source of concern and the severity of its incompleteness. The source of concern may differ from incompletely specified inputs, constraints, and internal model structures, to financial costs. The severity of the incompleteness can affect the explanation needs.
- **Time Limitation:** eventually interpretability is dependent on the way a model is used, therefore, the time that a user needs or is allowed to gain an understanding of an explanation is an important aspect. In certain domains where user time availability is restricted simple explanations are a necessity and in situations where the decision time is not a constraint a more complex and exhaustive explanation might be preferred. Another time limitation might be on the time it takes to produce the explanations. This depends on the computational complexity, considering technical explainability techniques, or when producing textual explanations rather exhaustively this could also take up a lot of time.
- **Nature of User Expertise:** users of models might have different experiences and domain knowledge and this has an impact on the level of complexity of an explanation. Therefore, a key aspect in defining the type of interpretability is to know the user experience with the model and task at hand. Experienced users may prefer a more complex elaborate model over a simplified model.

4.2.4. Good Explanations

Explaining and understanding are two different actions. Explaining depends on what is explained and how it is explained, this concerns the model and the interpretability method (Adadi and Berrada, 2018). On the other hand, understanding depends, in addition to these first two elements, on who is receiving the explanations. A ML model is only explainable when it is human-understandable and produces explanations to the intended user. Research from Miller (2019) articulates the link between Social Sciences and explainability by focusing on how to produce explanations that stimulate the human cognitive process (Adadi and Berrada, 2018). The most important point from Miller (2019) is that the explanation is contextual and not just the presentation of associations and causes (causal attribution). Explanations should consist of a small subset of relevant factors in which the recipient (audience) is interested, the explainer should select this subset, and interaction between both of them can lead to an optimization of this subset. A good, or human-friendly, explanation defined by Miller (2019) has the following properties:

- **Selective:** explanations can not cover the entire set of causes of an event. In order to comprehend an explanation, a selection should be made from the variety of causes and only a few must be chosen as the most important explanations. Such selectivity is deeply grounded within society as people often describe phenomena with one or two causal factors and do not include the entire set of causal factors. This also stems from the fact that humans can handle at most 7±2 cognitive entities at once and will not be able to process the entire list of causes of an event (Miller, 1956).

- **Contrastive:** humans are interested in why a certain prediction is made instead of another prediction, and not necessarily in the exhaustive causes of the current prediction. Also, contrastive explanations are easier to understand than complete explanations (Molnar, 2020). Thinking in counterfactual cases focuses on what a prediction would have been if input X changes. An example is that whenever a model predicts a rejection, e.g. in a mortgage application system, the first question that comes to mind is often focused on what factors need to change to get accepted. In such cases, people are interested in the contrast between the current input and the accepted version of their input. Contrasting explanations are application-dependent and need to have a point of reference in order to compare, this reference point depends on the data to be explained and on the audience. Molnar (2020) states that the best explanation is the one highlighting the greatest difference between the object of interest and the reference object.

- **Social:** explanations are social processes and are part of an interaction between the explainer and the audience. The nature and content of explanations are dependent on the social context and the environment in which the model operates together with the target audience. The content of explanations adapts to the social context, where for example explanations to fellow domain experts differ a lot from explanations to laypersons, implying that the explainer must be able to leverage the mental model of the audience while explaining (Adadi and Berrada, 2018).

When using explainability methods there are also technical properties of individual explanations that are desired from the explanations methods. Examples of properties of individual explanations are *accuracy*, *fidelity*, *comprehensibility*, and *consistency*. These properties are not in the scope of this thesis but should be considered, for an elaborate description of the individual properties of explanations produces by explanation methods Kaya (2022) should be visited.

4.2.5. Properties of Explanation Methods

Explanations can be either developed by humans or by methods that artificially generate explanations often using an algorithm. Earlier, the properties of good explanations are described in Section 4.2.4 and now the properties of explanation methods will be discussed. These properties are widely used to judge how good an explanation method is and are initially proposed by Lughofer et al. (2017) and Ribeiro et al. (2016), while extended by Robnik-Šikonja and Bohanec (2018). The properties of explanation methods are as follows:

- **Expressive Power:** is the language or structure of the explanation generated by the explanation method. This can vary from propositional logic (i.e. if-then rules), decision trees, and a weighted sum, to limited forms of natural language. The expressive power influences the comprehensibility of the explanation method and can be seen as the interpretability of the method itself.

- **Translucency:** describes the degree to which an explanation method relies on looking into the ML model and needs to know internal properties such as its parameters. Methods that have zero translucency do not rely at all on the model's internal structure or properties and may only use the input and observed predictions for example. High translucent methods rely on more information from the internals of the model to generate explanations and are in most cases model-specific.

- **Portability:** describes the ability of a method to cover the range of different ML models, in case methods are limited to neural networks their portability is low. Whenever methods have low translucency, not accessing model internals, they are highly portable and the methods with the highest portability are surrogate models.

- **Algorithmic Complexity:** describes the computational complexity of the method, specifically the underlying algorithm, that generates the explanation.

4.3. Explainable Artificial Intelligence

The definition of XAI or Explainable ML from Gianfagna and Di Cecco (2021) will be used and is as follows: *XAI is a set of methods and tools that can be applied to make ML models understandable to human beings in terms of providing explanations on the results provided by the ML models elaboration*. XAI is a method to achieve increased explainability, or interpretability, by applying techniques on the model, data, or environment. The discipline of XAI tries to make existing ML models more interpretable and understandable to humans, whether these models are black-box models or intrinsically interpretable models. There are many existing methods and tools for developing understanding which differ based on the necessary scope of interpretability or the type of models used. This has implications on the type of XAI tools that can be applied, this is determined by two choices. First, does the model allow for a *model-agnostic approach* or a *model-dependent approach* and the second question is on the scope of interpretability which must determine whether global or local explanations are needed.

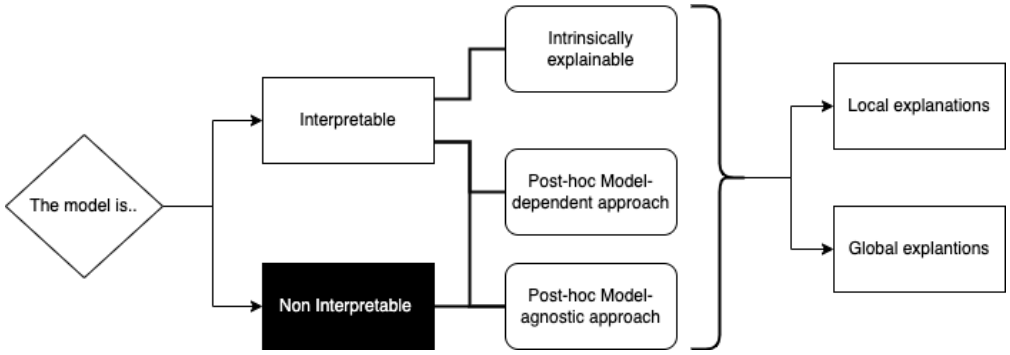


Figure 4.3.: Overview of Explainable AI approaches

4.3.1. Interpretable Models

The easiest way to ensure interpretability is to use the subset of algorithms that create interpretable models. Interpretable models fall in the set of ML models that can be understood by looking at their internal parameters, meaning that humans can eventually understand the cause of an output. In order to determine the rationale behind a model's decision the decision rules can be extracted or other model parameters can provide explanations directly. Interpretable models allow for direct interpretation of the models' internals when the model is simple, however, when the model becomes more complex the internals can be directly accessed to produce explanations. This property ensures that the models stand for their own explanations and that their explanations reflect perfect fidelity. The scope of interpretability depends on the complexity of the model but often falls within the level of global modular interpretability. When analyzing intrinsically interpretable models often three properties are considered which are: *linearity*, *monoticity*, and *interaction*. A model is linear when the association between the features and the target, the value to be predicted, is modeled linearly. Whenever the relationship between the features and the target outcome is always in one direction over the entire range of the features, the model has monotonicity constraints. This is when an increase in the feature value always leads to either an increase in the target outcome or a decrease. Monotonicity can be useful for the interpretation of a model because it will be easier to understand relationships between the features and the target outcome. Finally, the interaction between features can help in modeling more complex relationships and improve predictive performance, but too many or too complex interactions can dramatically decrease the interpretability of models (Molnar, 2020). These three properties can also be forced upon models, whenever the internal structure is accessible. An example of this can be to force relations within nonlinear models to be monotonic so that the direction of the change in outcome is always the same, this can be used to increase explainability.

The three main behaviors of the learned function in interpretable models can give an indication of the level of transparency. The level of transparency of interpretable models has been formalized within the research from Barredo Arrieta et al. (2020) and is algorithmic transparency, decomposability, and simulatability. The three levels of transparency will be elaborated down below:

- **Simulatability:** is the ability of a model of being simulated or comprehended by a human. Complexity is the main factor influencing simulatability as even simple models which

become too large can result in the inability of humans to think and reason about it as a whole. Predictors should be human-readable and the interactions between them should be kept to a minimum.

- **Decomposability:** is the ability of a model of being able to explain each part, this characteristic empowers the global interpretability of a model resulting in understanding, interpreting, and explaining the behavior of a model. This requires every input to be understandable by humans without using additional tools, the variables should still be understandable but the predictors and interactions can be of such a size that decomposability is needed.
- **Algorithmic Transparency:** is the ability of a model to be able to understand the processes it follows to produce the output from its input data. This allows users to understand and reasons about how the model acts in different kind of situations, this implies that the model must be entirely explorable by means of mathematical analysis and methods because the variables and interactions are too complex to be analyzed without them.

The models that fall in one or all three levels of model transparency form the suite of interpretable models and are Linear Regression, Logistic Regression, Decision Trees, RuleFit, Naive Bayes, and k-nearest neighbors. The most well-known interpretable models are Linear Regression, Logistic Regression, and Decision Trees and will be taken as representatives for discussing the levels of transparency, these models and their transparency are further discussed in Appendix I.

4.3.2. Blackbox Models

Blackbox models are ML models that are too complicated for any human to understand their inner workings. The term black box is often also used whenever models are proprietary and the model cannot be inspected by third parties (Rudin et al., 2022). Guidotti et al. (2018) defines a black box predictor as "*a data-mining and machine-learning obscure model, whose internals are either unknown to the observer or they are known but uninterpretable by humans*". Blackbox models entail all models whose internal workings are uninterpretable by humans often due to complexity. Techniques such as Deep Learning (DL) are most often considered to produce black box models but even complex interpretable models, such as decision trees, can become uninterpretable and blackboxes. Guidotti et al. (2018) categorizes the problems coming forth out of opening up the black box into three different problems: model explanation, outcome explanation, and model inspection.

4.3.3. Model-dependent Approach

Model-dependent methods are built specifically for the model to be explained and rely on accessing the model's internal structure and parameters (Gianfagna and Di Cecco, 2021). This results in the fact that a particular type of interpretation is necessary depending on the specific model class limiting the choice to potentially apply more predictive and representative models (Adadi and Berrada, 2018). Model-specific explanations are usually complex because they deal with the internal structure of the model, hence a certain level of expertise is required in order to interpret them. Therefore, there has been a growing interest in model-agnostic interpretability methods that are not limited to working with a specific set of ML models (Adadi and Berrada,

2018). Because model-dependent methods are so specific they will not be addressed singularly within this research.

4.3.4. Model-agnostic Approach

Model-agnostic methods do not access the internal parameters of the model to be explained. This makes these methods extremely portable with low translucency as they are not tied to a particular type of ML model. Agnostic methods have the advantage of being extremely friendly to users and can be used in order to understand or test models without knowing any prior information about the methodology used to train the model (Gianfagna and Di Cecco, 2021). Because the model-agnostic methods do not rely on the internal workings of the model they can separate prediction from explanations, making them easily transferable for users on multiple models which can lead to a growing understanding of working with the interpretability method. Model-agnostic interpretations are usually post-hoc, referring to the fact that they are used to interpret an event that has occurred (the prediction). Post-hoc techniques are analogous to the human way of explaining decisions as the interpretability of humans is often post-hoc itself (Adadi and Berrada, 2018; Lipton, 2018). In the past years, a large amount of model-agnostic methods has been developed which can be categorized based on the four technique types: Visualization, Knowledge extraction, Influence methods, and Example-based explanation (Adadi and Berrada, 2018). The technique types and well-known interpretability methods that fall into these classes are listed in Table 4.1. The interpretability methods that are of relevance for the scope of this research are Surrogate Models, Feature Importance, and Counterfactual Explanations. The Surrogate Model interpretability method Local Interpretable Model-agnostic Explanations (LIME) will be discussed together with the Counterfactual Explanation method in Appendix J. Because Influence Methods techniques and particularly Feature Importance methods are the most used in practices these will be further discussed below. An overview of all the technique types and interpretability methods defined by Adadi and Berrada (2018) are given in Table 4.1.

Technique type	**Interpretability method**
Visualization	Surrogate Models Partial Dependency Plot (PDP) Individual Conditional Expectation (ICE)
Knowledge Extraction	Rule Extraction Model Distillation
Influence Methods	Sensitivity Analysis Layer-wise Relevance Propagation Feature Importance
Example-based Explanations	Prototypes and Criticisms Counterfactuals Explanations

Table 4.1.: Overview of model-agnostic methods categorized by the four technique (Adadi and Berrada, 2018)

Influence Methods

This type of technique estimates the importance or relevance of a feature by measuring the change in model performance whenever the input or internal components change. Influence

techniques can also be visualized increasing the understandability to users. One of the most well-known interpretability methods within the set influence method techniques is feature importance which shows the contribution of features toward the prediction of the underlying model. Feature importance methods are widely used in practice and also within TM. One of the most applied feature importance methods, SHapley Additive exPlanation (SHAP), will be described below.

Feature Importance - Shapley
The usage of Shapley values is perhaps one of the most applied and famous XAI methods and stems from coalition game theory. The Shapley method is a local interpretability method that explains predictions by assuming that each feature value of the instance is a 'player' in a game. Within this game, the prediction is the payout and the method will generate a fair distribution of the payout among the features. This game resembles to which degree a feature is responsible, or deserves the payout, towards a certain prediction. Eventually, Shapley values are generated for each of the features resembling the feature contribution towards the prediction. A solid effort to leverage Shapley values has resulted in SHAP, developed by Jalali and Wohlin (2012) and is the most renowned application of Shapley values for ML. Contrary to LIME, SHAP is the only method to deliver a complete attribution and in addition to this, it allows for contrastive explanations resulting in the ability to compare predictions to a subset or a single data point. Shapley is also the only explanation method with a solid theory whereas the explanations adhere to the axioms efficiency, symmetry, dummy, and additivity. To read more about these axioms the book *Interpretable Machine Learning* from Molnar (2020) is recommended.

The main disadvantage of SHAP is that it explains the feature correlation defined by the model but does not imply causality. Therefore, to properly use SHAP it is important to determine and examine the features in the model for importance, signage, and causal behavior. Next to this, SHAP can have limitations on the computing time, when searching for an exact computation of the Shapley value there are $2^{feature}$ possible combinations of the feature values. There are possibilities to decrease the computation time by sampling combinations and limiting the number of iterations, however, this increases the variance of the Shapley value. Another disadvantage is that Shapley values are often misinterpreted as they are often seen as the difference of the predicted value after removing the feature, this is not correct. The Shapley value represents the contributions of a feature value to the difference between the actual prediction and the mean prediction, given the current set of feature values (Molnar, 2020). In order to calculate the Shapley value for new data instances the original data needs to be accessed which can lead to problems when the dataset is unavailable. This is represented in SHAP as it is highly dependent on the model, SHAP shows the feature importance to the model and not the feature importance in reality. Finally, the Shapley value method does not work perfectly with correlated features because of the fact that the interaction between features can not be taken into account through the game theory-based approach.

4.4. Revisiting the View on Explainability

Explanations are a form of social interaction having psychological, cognitive, and philosophical projections but still, the ideas of Social Sciences and human behavior are not yet visible in the field of XAI (Adadi and Berrada, 2018). Current explainability approaches rather focus on the technical aspects of ML models and are based on the intuition of the developer and less on

Property	Assessment
Expressive Power	High, for single predictions.
Translucency	Low, no insights into the model internals.
Portability	High, the method does not rely on the inner workings of the ML model.
Algorithmic Complexity	High, computation time can increase significantly and harder to get a full understanding.

Table 4.2.: Properties of SHAP explanation method

the audience or intended user (Barredo Arrieta et al., 2020; Miller, 2019). Besides this, questions may be raised about whether explainability should be solved with technical approaches or by designing inherent interpretable ML models. In addition to this, the type of explanations differs based on environmental latent dimensions such as time and the domain in which the ML model is deployed. Therefore, this research proposes to include these social and organizational elements in approaching explainability by taking the audience into account and using a socio-technical perspective.

4.4.1. Focus on the intended Audience

XAI encompasses methods, procedures, and strategies to provide explanatory information helping to understand systems that are too complex for human oversight or are inherently opaque (Langer et al., 2021). The need to understand stems from the different human stakeholders each having their own interests, goals, expectations, needs, and desires regarding the ML system. However, current approaches of XAI are focused on developing new XAI approaches without taking into account the desiderata of the intended user and stakeholders. This shows the mismatch between current research and the eventual objective of XAI as the success of XAI approaches depends on the fulfillment of the stakeholder desiderata. There are current measures and metrics focusing on how well approaches calibrate trust or improve human-machine performance but these are often only translated into technical concepts.

Within AI systems there are multiple types of stakeholders, research from Tomsett et al. (2018) proposes seven roles involved in the creation, usage, and maintenance of AI systems:

- **Creator-Owner:** represents the owner of the intellectual property in the AI system, these are often business managers or executive board members.
- **Creator-Implementer:** is the direct implementer of the AI system, these are often data scientists, developers, product owners, and domain experts.
- **Operator:** gives input to the AI system and retrieves its predictions, these are often domain experts and/or users of the model.
- **Executor:** makes decisions based on the AI system's predictions, these are domain experts and/or users of the model. Decision subject
- **Decision Subject:** is the entity affected by a decision based on the prediction of the AI system.
- **Data Subject:** is the person whose personal data has been used to train the AI system.

- **Examiner:** audits and investigates the AI system, these are often regulatory entities or corporate auditors.

All involved stakeholders in AI systems might need explanations on the data used, global model behavior, or local model predictions. The needs for explanations can differ based on these stakeholder roles which influences their desiderata. The desiderata of stakeholders contain the reason for explainability that empower the functioning of their role in the system. Next to this, the desiderata can vary per stakeholder role due to their prior knowledge and understanding of the system and techniques applied, this is represented within the mental model of the stakeholders. The mental model of the stakeholder represents the level of expertise, domain knowledge, cultural background, interests and preference, and other contextual variables (Adadi and Berrada, 2018). Next to this, they can have various reasons motivating their needs and requests of explainability, earlier pictured in Figure 4.1. Both these aspects influence the desiderata of the stakeholders, in order to satisfy these desiderata understanding is required. Understanding requires explanatory information which can be fulfilled by explainability approaches, however, explainability approaches are limited to the system environment they are applied in and the technical constraints coming forth out of the technical design choices.

There is a growing need for defining the desiderata of the intended user. Research from Barredo Arrieta et al. (2020) even includes the intended user in defining XAI which goes as follows: *"Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand"*. This newly defined definition is supported throughout this research and will be applied by performing systematic empirical research. This analysis will define the involved stakeholders in the process of TM and investigate their desiderata and reasons for explainability in Chapter 6.

4.4.2. Explainability From a Socio-technical Point of View

Apart from taking into account the involved stakeholders and their desiderata it is important to include organizational latent dimensions of explainability. Next to this, general organizational elements should be considered to design and especially maintain explainability within the ML system. When ML models are applied in practice there are often multiple points of interaction between components in the system. Current research on XAI is often only focused on the interaction between the developer, explainability technique, and the ML model. However, there are other factors to be considered when designing for explainability. Realizing explainability should be addressed by multi-disciplinary efforts due to the nature of explainability being a multifaceted objective (Adadi and Berrada, 2018). Explainability can only be achieved when it is taken into account within multiple facets and layers of the system, hence, a shift should take place from technosolutionisms to taking on a socio-technical perspective. This Chapter has defined multiple factors influencing explainability which can be categorized within the socio-technical system, the combination of these elements influencing explainability are displayed in Figure 4.4. Approaching explainability should be done by designing explainability practices that include, next to the audience, the hardware, software, and organizational factors of the entire socio-technical system.

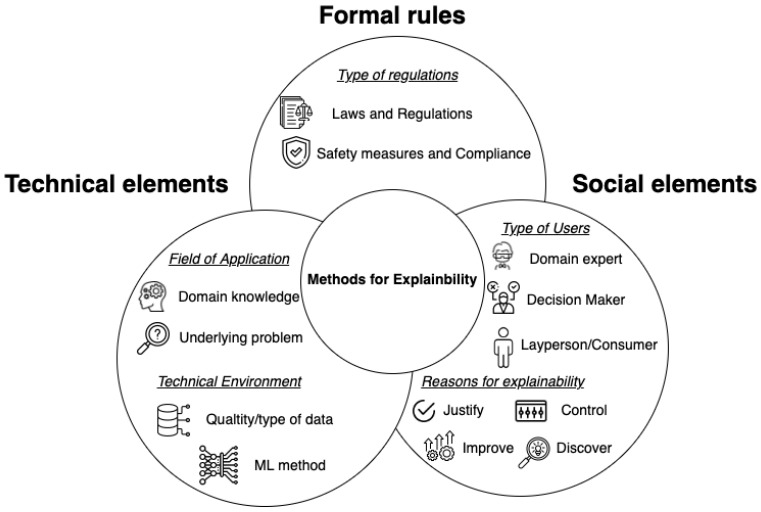


Figure 4.4.: Influencing factors of explainability

4.5. Main Findings Chapter 4

This Chapter tries to answer research sub-question two and three and the results will be discussed down below.

SQ2: *What are the important elements necessary to define good explainability?*

Explainability can be defined as the ability to explain or present in understandable terms to a human. Explainability can stem from four main reasons which are to justify, control, improve, and discover. Each reason might need a different scope of explainability fit for the application at hand, the scope of interpretability can vary from high-level algorithmic transparency to global model explainability and local explainability. Explainability is stemming from the reasons that can empower stakeholders in their activities and can be categorized as reasons to justify, control, improve, and discover. But explainability is also influenced by latent dimensions such as the domain the ML model is deployed in, the severity of incomplete or faulty explanations, and the time available to receive and interpret the explanations. There have been studies from Social Sciences defining good explanations stating that they must be selective, contrastive, and social. However, technical explanations produced by explanation methods can be assessed by elements such as the accuracy, fidelity, comprehensibility, and consistency of the explanation. Explanation methods also have properties that are categorized as the expressive power they produce, their translucency, portability, and the algorithmic complexity of the method. All the elements above constitute toward the definition of good explanations and good explainability methods while showing the complexity and dependence of explanations on the intended user and other latent dimensions.

SQ3: *What are the limitations and challenges of explainability practices applicable to TM?*

Within research, there are multiple debates on XAI and explainability starting with the lack of consensus on definitions such as explainability and interpretability. Also, there is still ambiguity on the classification of explainability methods and ways to assess or evaluate them. There

has been an enormous amount of research on XAI especially focusing on model-agnostic methods usually being post-hoc. These type of methods all have their technical limitations and discrepancies. However, the main limitation is the fact that explainability is a social process and is approached by technocentric solutions. There has been limited research in applying a user-centered perspective or taking into account organizational elements that highly influence explainability practices and needs. Therefore, this research proposes to include the audience in approaching explainability and even to extend this by stating that explainability must be viewed from a socio-technical perspective. This is due to the fact that explainability is dependent on the type of business, field of application, technical environment, the type of user, and the reasons for explainability. Currently, most efforts are focused on the technical environment which is extremely limiting. Next to this, there are not yet any practical frameworks that can help the local practice to apply the view of a socio-technical approach to explainability. There is some existing research on XAI as a socio-technical problem that is suggesting strategies, for example, research from de Bruijn et al. (2021) proposes to shift the focus from explainable algorithms to explainable processes. This research will build upon such strategies by incorporating a socio-technical perspective on explainability and will try to develop a structured approach that can be applied within the practical environment of TM.

5. System Safety Engineering

This Chapter explores systems theory and system safety theory to determine what concepts and strategies can be used in approaching explainability. The Chapter is within the second phase of the research contributing toward the knowledge base. This Chapter will try to answer the fourth research sub-questions, which is as follows:

SQ4: *What concepts and strategies of system safety theory can aid explainability practices?*

5.1. Systems Theory

System theory lays the foundation to understand and engineer complex systems. Generally speaking, systems can be divided into three categories, those that exhibit *organized complexity*, *unorganized complexity*, and *organized complexity* (Weinberg, 2001). The systems that exhibit organized complexity are often too complex for complete analysis and too organized to approach them with statistics (Leveson, 2011). These systems are often social systems, biological systems, and software-based systems, but also those that intuitively seem less complex. System theory is developed for organized complex systems and focuses on systems as a whole assuming that some properties of systems can only be analyzed in their entirety, including aspects ranging from the social to the institutional and technical levels. Concentrating on the design and analysis of systems exhibiting organized complexity, system theory is based upon two main pairs of ideas: (1) *emergence* and *hierarchy* and (2) *communication* and *control* (Ashby, 1957; Leveson, 2011).

5.1.1. Emergence and Hierarchy

Emergence is the phenomenon that is described by Bouwmans (2020) as:

"when behavior at a higher level of aggregation in a system results from the behavior of the constituting parts at a lower level of aggregation".

Complex systems are often expressed in terms of a hierarchy of levels of the organization where complexity increases from the lower levels upwards (Leveson, 2011). Constraints or the lack of constraints on components, or the potential interaction between components, at higher hierarchical levels, allow lower-level behavior (Leveson, 2011). Within these systems, emergent properties can arise throughout higher levels depending on the enforcement of these constraints. The emergent properties arise through interacting processes within the lower levels, thus becoming more complex when moving up levels of abstraction or hierarchy. These emergent properties do not exist at the lower levels but are created by the interacting processes that in isolation may seem simple and do not exhibit such properties. Emergent properties can only be determined at a system level, taking into account the system as a whole.

In determining whether systems are safe it should therefore be acknowledged that safety can only be determined by taking into account the system as a whole, defining safety as an emergent property. When considering the safety of single, isolated, components the interactions and behavior in higher levels of the hierarchy are not considered resulting in the inability to call the component or sub-system safe. Systems theory tries to include the interaction of components and processes, higher-level complexity, and emergent properties.

5.1.2. Communication and Control

The second main idea laying the foundation of systems theory is that of *communication* and *control*. Control can be enforced by imposing constraints on certain levels of a hierarchy. The control processes, defining the 'laws of behavior', operate at the interface of different levels in systems and are defining the hierarchy between them (Checkland, 1981; Leveson, 2011). Systems that are influenced by their environment, called open systems, are in need of processes of *communication* keeping them in a state of dynamic equilibrium and being able to adapt to changes (Ashby, 1957). These communication processes, in which information is communicated, involve both the system together with their interrelated components and the environment (Checkland, 2012; Leveson, 2011). Often feedback loops are used as a process of communication enforcing regulation or control. An example of communication channels in TM between control levels is given in Figure 5.1. The downward *reference channel* provides information on given system constraints and the upward *measuring channel* can provide operational feedback on how the constraints are being satisfied (Leveson, 2011).

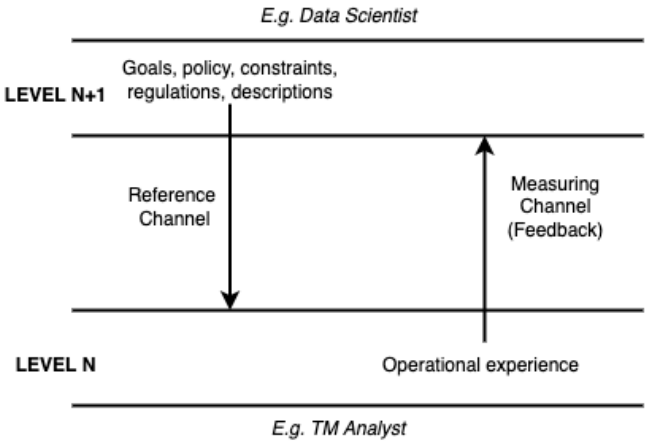


Figure 5.1.: Communication channels between control levels from Leveson (2011).

In the example of ensuring safety in systems, control processes can consist of imposing constraints on system behavior to avoid unsafe events or conditions (Leveson, 2011). When controlling a process the entity controlling the process is referred to as the controller, the controller can either be a control algorithm, physical machinery, or even a human operator. There are four conditions, described by Ashby (1957), to control a process and are captured in a process model. The four conditions are:

1. *Goal Condition*: The controller must have a goal, which is the constraint that must be enforced.

2. *Action Condition*: The controller must be able to affect or influence the system state, this can be done in the downward reference control channels.
3. *Observability Condition*: The controller must be able to determine the current state of the system, this is embodied in the upward feedback or measuring channels.
4. *Model Condition*: The controller must have a model of the process being controlled in order to control it effectively.

5.2. System Safety Engineering

The origin of safety engineering goes back centuries, however, defined structured approaches for designing safe products and systems arose in the postwar era. During this time period, societal concerns and public debate opened up conversations on the safety of nuclear power, civil aviation, the development of lethal chemicals and weapons, and increased environmental pollution (Leveson, 2003). There was a growing need for ensuring safety for systems that can cause hazards such as human loss or injury but also the destruction of property and environmental harm (Leveson, 2003). The classical approach for safety engineering was based on analytical reduction, isolating system components, and applied to event-based accident models. There have been growing accidents concerned with systemic factors such as the lack of proper procedures, guidance, and training which traditional safety engineering do not consider. Appendix K provides further information on the history and methods used in traditional safety engineering.

The traditional approach does not longer hold for the large and complex systems that are being built today. According to Leveson (2011) the main changes, amongst others, that stretch the limits of traditional safety engineering is the fast pace of technological change, increased complexity and coupling, more complex relationships between humans and automation, changing regulatory and public views of safety, new types of hazards, decreasing tolerance for single accidents, and the changing nature of accidents. Leveson (2011) states that a paradigm shift is needed to extend the understanding of accident causation and prevention techniques and proposes a system safety engineering approach that tries to facilitate this change by incorporating systems theory and systems thinking in the traditional safety engineering approach. Within the book *Engineering a safer world: systems thinking applied to safety*, Nancy Leveson develops this new approach to safety by applying systems theory to safety engineering. Within her work, seven traditional assumptions of safety engineering are stated and evaluated by their relevance to today's complex systems. The assumptions are reconsidered and put into perspective by using this new approach to system safety engineering. This resulted in newly proposed assumptions on approaching safety developed by Leveson (2011) and are coined by Dobbe (2022) as *Leveson Lessons*. These lessons are pictured in Appendix M Figure M.1. For an elaborate description of the *Leveson Lessons* Appendix L can be visited.

System safety engineering is renewing the traditional safety engineering efforts by introducing systems theory and viewing safety as an emergent system property that must be controlled for. System safety engineering uses systems theory and systems engineering and starts in the earliest concept development stages of a project and remains relevant within the subsequent design, production, testing, operational use, and disposal phases (Leveson, 2003). System safety is defined by Leveson (2003) as:

"*a planned, disciplined, and systematic approach to identifying, analyzing, and controlling hazards throughout the life cycle of a system in order to prevent or reduce accidents*".

System safety engineering, based on systems theory, treats safety as an emergent property considering that accidents can arise from the interactions among system components or with the environment and does not necessarily seek for a root cause or causality (Leveson, 2002). This instantiates safety as a control problem whereas accidents result in uncontrolled interactions or violation of the set of constraints. System accidents are now viewed as phenomena when components fail, external disturbances occur, dysfunctional interactions between components happen, safety constraints are inadequately enforced, or a combination of all occurs. System safety engineering provides an improved approach to traditional safety engineering which allows for designing safety in the system as it is being (re)developed. As complex socio-technical systems become larger and complexity increases the call for good system engineering approaches becomes more critical in order to guarantee a safe deployment into society. Incorporating system engineering processes for approaching safety combined with a systems theory approach may establish effective means for the design of resilient and safe complex systems as a whole. System safety engineering can guide the adaptation of technical advancements in many fields and is concerned with preventing foreseeable accidents while trying to minimize the result of unforeseen ones (Leveson, 2003). These practices lead to the ability to successfully engineer safer and more complex systems that allow for interactive complexity and coupling, resulting in the eventual objective of system safety engineering. One of the leading approaches for ensuring safety is the system safety engineering approach from Leveson (2003), focusing on safe management, development, and operations of socio-technical systems in society.

5.2.1. Safety as a Control Problem

System safety engineering has reintroduced safety as a control problem that can be enforced by behavioral safety constraints. In addition to this, accidents are no longer merely analyzed as single-component failures but are extended to incorporating interactions between components and the environment. In order to approach safety Leveson (2011) has developed a framework that applies a system-theoretic view on causality. This new expanded accident causality model incorporates the Leveson Lessons and is called Systems-Theoretic Accident Model and Processes (STAMP). The originating concepts of STAMP are derived from the four properties of systems, described by Checkland (1981) as hierarchy, emergence, processes of communication, and control processes. The three basic constructs underlying STAMP are safety constraints, hierarchical control structures, and process models.

STAMP can aid safety-guided design by identifying the constraints required to maintain safety. It can do so by first trying to identify flaws existent in the control structure which tries to enforce the safety constraints in order to prevent accidents from happening. Then, the control structure can be (re)designed next to the physical system and operating conditions that enforce the constraints. Often during this process, new safety constraints can be designed and implemented. The following subsections will elaborate on the three main constructs of STAMP.

5.2.2. Safety Constraints

Losses, or accidents, occur due to the violation or lack of safety constraints. Constraints are limitations on the behavioral degree of freedom of the system components and are often focused on *how* a system reaches its goal (Leveson and Stephanopoulos, 2013). With the complexity of systems increasing it has become more difficult to identify and enforce safety constraints in the

design and operations (Leveson, 2011). The complex systems of today are most often software-based and are not limited by physical and operational constraints (Leveson, 2011). In these new systems, there are almost no limits on complexity which makes the design more difficult due to the absence of these 'natural' constraints imposed previously on physical systems.

In order to ensure safety at a system-level safety constraints must be identified in order to design appropriate control processes to enforce them. Constraints can be broken down into sub-requirements or constraints allocated to certain components of the system design. Clear responsibility must be allocated to involved actors for the control processes that enforce the identified safety constraints. Trying to enforce these constraints will lead to the establishment of design decisions focused on safety in the systems, this shows the importance of identifying constraints early in the development.

5.2.3. Hierarchical Safety Control Structure

Systems theory views systems as hierarchical structures where behavior within levels can be controlled. Constraints at higher levels can allow or control lower-level behavior, to integrate these constraints it is important to instantiate the control processes between the levels. Each control process is responsible for certain safety constraints. Accidents happen when the control processes provide inadequate control which can lead to a violation of the safety constraints. Designing hierarchical safety control structures can be very difficult because when constraints are missing at the lower level this will allow for unassigned responsibility for safety and can lead to hazardous states of the entire system.

Between each level of the safety control structure communication channels are needed to communicate the safety constraints to the lower level and receive feedback on the operational experience checking how effectively the safety constraints are being satisfied. An example of such a communication channel is provided earlier in Figure 5.1. Because control structures are adaptive and change over time it is extremely important to evaluate and analyze them to determine whether the safety constraints can still be effectively managed by the control processes. When analyzing hazards not all hierarchical levels are relevant but it is important that they are identified at a system level, after this, a top-down approach can identify the safety constraints for the parts of the overall control structure (Leveson, 2011). A thorough analysis of the hierarchical safety control structure can lead to new procedures or new control processes that will ensure the defined safety constraints. Figure 5.2 shows a generalization of the socio-technical hierarchical safety control structure applied to the TM system. It can be seen that often structure consists of two hierarchical control structures that have interactions with each other, one for development and another for operations. This shows to which extent control can be enforced. Within the structures, it shows that safety during operations is dependent on the original design and the operating procedures and that there is a need for communication channels between them (Leveson, 2011).

5.2.4. Process Model

The process model is constructed of the four conditions that are required to control a process which is described in Section 5.1.2. The process model is a model the controller has of the process it is controlling, only when the controller has such a model it can provide adequate control

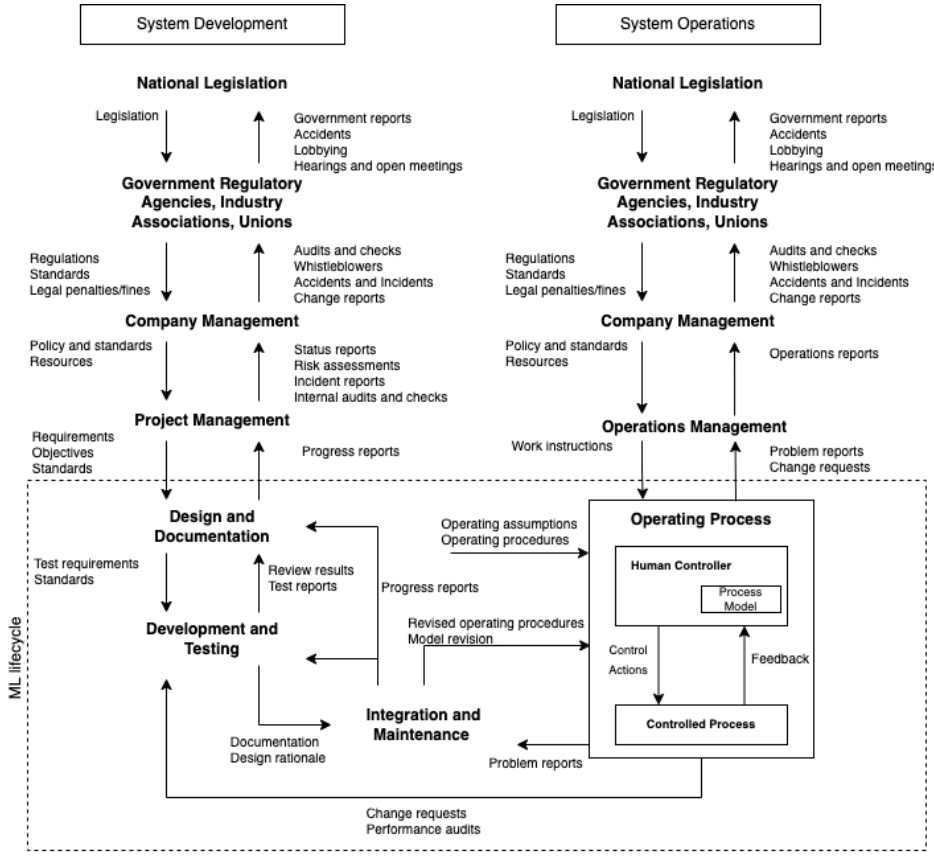


Figure 5.2.: Model of socio-technical control of transaction monitoring system

instructions. The process model is only applicable to the controller that is active within the operating process. A typical control structure is provided in Figure 5.3, where it can be seen that the controller (either human or automated) uses the process model and the decision-generating process (or control algorithm if the controller is automated) to generate control actions on the controlled process. And the controlled process can then again update the process model to the current state of the process. There could be multiple reasons for incorrect process models and most have to do with the feedback function. Whenever there is no feedback, it is not received, it is delayed, or it is incorrect this can lead to incorrect process models of the current state of the process. In such cases, the controller has a wrong process model and can make faulty assumptions leading to undesired scenarios.

Hence, dysfunctional interactions can be explained in terms of incorrect process models. Such component interaction accidents, involving complex digital technology or human error, often occur when the actual process does not match the process model used by the controller. When this mismatch is existent it can lead to inadequate control actions which are formally described by Leveson (2011) as:

1. Required control actions (for safety) are not provided.
2. Incorrect or unsafe control commands are given.

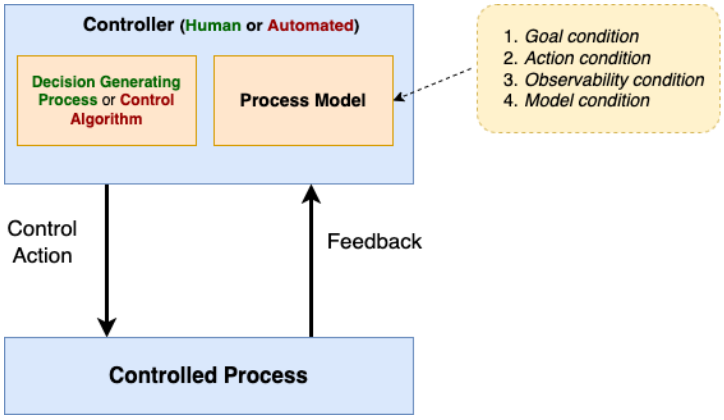


Figure 5.3.: Control process related to process model, inspired by Leveson (2011)

3. Potentially correct control commands are provided at the wrong time, too early, or too late.
4. Control is stopped too soon or applied too long.

Analyzing the process model can generate potential new design features and requirements and can be used during development and operations. Process models can play an important role in finding out why accidents happened and how inadequate control is established by the human or operator, also, they can help in the design of safer systems by effectively investigating and controlling component interactions (Leveson, 2011).

5.3. System Safety for Artificial Intelligence Systems

Computers have become so powerful and useful that they have eliminated many of the constraints that previous physical systems had to deal with, this is both a blessing and a curse. Hence, the *curse of flexibility* is the main problem within software-based and AI systems. Due to the absence of physical constraints, there is (almost) nothing limiting the complexity of the design. The question with software-based systems is not focused on what is possible, but rather on what can be accomplished successfully and safely. Often software, or AI models, go beyond the human intellectual limits of understanding. Research from Leveson (2011) states that:

"safety-related software errors most of the time arise from (1) discrepancies between the documented requirements specification and the requirements needed for the correct functioning of the system and (2) misunderstanding about the software's interface with the rest of the system".

The accidents in software-based systems did not occur from coding errors but rather from what is called flawed requirements. This shows the importance of requirements specification and the most important element is to determine these requirements in order to specify what the software should do and what it not should do. Leveson (2011) stipulates the need for system safety, especially with the rise of new hazards that software and automation bring to systems. Hence, Leveson (2011) elaborates thoroughly on hazards caused by interactive complexity, misalignment of mental models, incomplete requirements, and the curse of flexibility. These are all, among others, highly relevant for software-based systems. Within society, there has been a

tremendous increase in software-based systems, especially those implementing AI replacing human analysis or decision-making within systems. Although AI have been around for decades, there has been a rapid increase in the development and deployment of AI systems in society. As AI systems are being incorporated within organizations they are integrated into existing systems and often function in a process that informs assists or automates human decision-making. Hence, AI systems can be viewed as socio-technical systems which consist of two subsystems: the social subsystem, which includes people and relation structures, and the technical subsystem that encompasses the technology, processes, and procedures, referring to the physical environment (Castro et al., 2020). As the adoption and implementation of AI systems have been increasing, concerns are being raised about the societal impacts of such AI systems. One of the main topics of discussion is the problem of accidents in AI systems. Within this scenario, accidents are defined as unintended and harmful behavior that may emerge from using AI systems (Amodei et al., 2016).

The work from Dobbe (2022) taps into this by focusing on preventing harm in AI systems based on the Leveson Lessons and formulates strategies to possibly prevent these harms. The core assumption is that safety can not be safeguarded merely through technical design choices and a system perspective should be applied taking into account the social, institutional, and technical components of a system. Establishing the view of a socio-technical system will approach values such as safety as emergent properties that can not be ensured within one particular level of a system but are a result of implementation throughout the system as a whole. Dobbe (2022) links the Leveson Lessons to the implications for AI system safety and suggests a strategy for each of them. The Leveson lessons, implications, and examples of specific countering strategies are pictured in Appendix M Figure M.1 and are taken from Dobbe (2022). These system safety strategies for AI systems, linked to Leveson Lessons, are elaborated down below.

Strategy 1: ***Identify hazards at the systems rather than components level.***
Dobbe (2022) suggests that in order to facilitate designers with the ability to translate identified hazards into concrete requirements the boundaries of the systems should be drawn including the conditions related to accidents over which system designers have some control. The concrete requirements can be in the form of constraints from technical capabilities or can be on operations, which are often constrained by organizational or regulatory bodies. This enables designers to design hazards out of the system and focuses on states the system should not be in (Dobbe, 2022). However, system developers can not have control over all conditions related to accidents. Some of these conditions come out of the institutional context and should be properly addressed in order to allocate responsibilities, within system safety the institutional design is referred to as the *safety control structure*. Dobbe (2022) portrayed the relationship between the AI system design, safety control structure, and the accident model which can be seen in Figure 5.4.

Strategy 2: ***Ensure safety through socio-technical constraints.***
The proposed systems-theoretic view uses constraints on the system's components and their interactions to ensure the system does not cross the boundary of a hazardous state. These constraints should be enforced and this is done by designing control structures, which are controlling for the constraints. Control structures can be designed for human operations, integrated into the physical design of the system, or through social forms of control such as formal rules, or even attained in the values and norms embedded in the organizational culture (Dobbe, 2022).

Strategy 3: ***Capture the safety condition and assumptions in a process model.***
In order to effectively design safety measures and socio-technical fail-safe mechanisms, the boundary of the system and reach of safety should be clearly defined (Dobbe, 2022; Dobbe

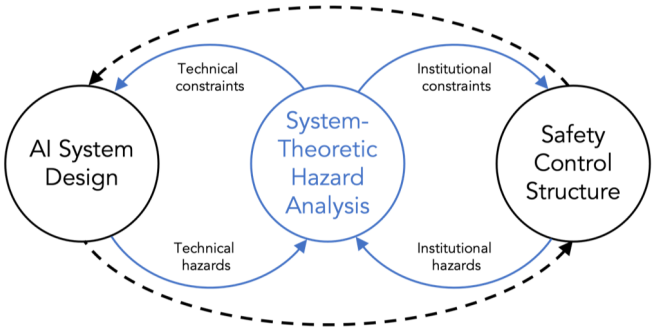


Figure 5.4.: Integration of accident models in the design of Artificial Intelligence systems, including the design of the institutional safety control structure. Taken from Dobbe (2022).

et al., 2021). The boundary of analysis should include the AI model and its interaction with the environment, taking into account other system components, human operators, users, and involved processes (Dobbe, 2022). In order to do so, a *process model* can be designed that will include these elements to structurally define the boundaries of the analysis. A process model can aid in specifying where potential fail-safe mechanisms could be designed to enforce the set of safety constraints and will foster a deeper understanding of how accidents occur (Dobbe, 2022). Good process models that match the necessary processes for controllers (automated or human) can effectively safeguard safety constraints and mitigate control errors.

Strategy 4: ***Align mental models across design, operation, and affected stakeholders.***
AI systems often help to inform, assist, or automate decisions for a human operator, this human-machine interaction can lead to a set of hazards. As earlier discussed, is that accident analysis should not merely focus on operator error, or deviation from normative procedures, but rather on the environment and information available to the operator. A mental model is an overarching term for any framework, worldview, or concept a person carries in their mind to be able to understand phenomena. Mental models are naturally evolving through interaction with the specified system and are constrained by a person's technical background, experience with similar systems, and ability to process information (Gentner and Stevens, 2014a). Mental models differ per individual and are based on generalization and analogies from experiences, but mental models can also be defined for certain roles when setting up certain general assumptions (Gentner and Stevens, 2014b). Figure 5.5, developed by Leveson (2002), shows the relationship between the actual system and the mental models of the designer and the operator.

Hence, it is important to align the mental models, depicted in Figure 5.5, between the designer and the operator to get a thorough understanding of the effective procedures to better safeguard and optimize the system through first-hand experience (Dobbe, 2022). The mental models used when interacting with the system emerge over time and should be periodically updated. In order to successfully prevent human error the designers should understand why and how choices are made so that appropriate controls can be included in both the design of the system and the environment. The work from Leveson (2011) outlines three principles for preventively taking into account possible human errors. These three principles are as follows:

1. Design for redundant paths: *to provide multiple paths to ensure that a single error cannot prevent the operator from taking action to maintain a safe system state and avoid hazards;*

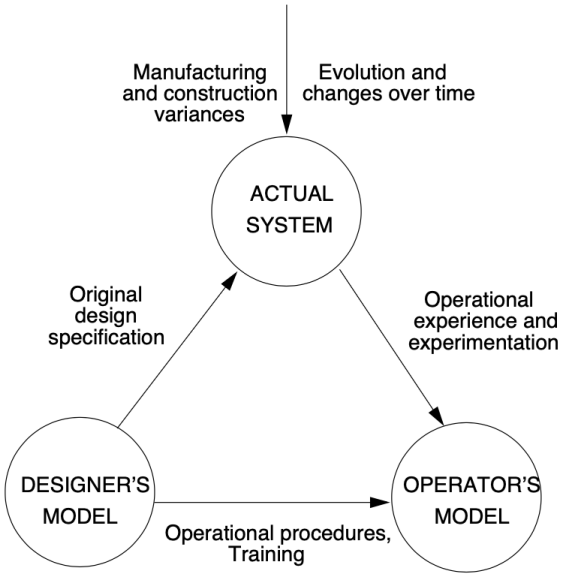


Figure 5.5.: The relationship between mental models from Leveson (2002).

2. Design for incremental control: *to give the operator enough time and feedback to perform control actions and, if possible, to do so in incremental steps rather than in one control action;*
3. Design for error tolerance: *to make sure that reversible errors are observable to the human operator before unacceptable consequences occur, to allow them to monitor their own performance, and to recover from erroneous actions.*

Strategy 5: ***Include software and related organizational and infrastructural dependencies in system-theoretic hazard analysis.***
The properties of software-based systems bring new challenges such as increased complexity and specific expert knowledge leading to a set of new hazards to the effort to understand or communicate newly arising system requirements. In order to enforce the safety constraints of these systems accident models should thoroughly analyze hazards and provide information and documentation (Dobbe, 2022). This information, coming from the accident model, should provide a base for setting the requirements of a software-based system used in the process model. Next to this, it can help adjust the institutional constraints accordingly in the safety control structure.

Strategy 6: ***Organize feedback mechanisms for operational safety.***
Because the migration of systems to a higher state of risk is predictable and manageable, anticipation can try to eliminate these hazards (Rasmussen, 1997). Such anticipation should be addressed by using operational safety control structures that enable control and feedback loops (Dobbe, 2022). Feedback mechanisms are used by the controller fed by the AI system in real-time to control processes (Dobbe, 2022). However, feedback mechanisms should also be applied on a system level to maintain and update operational safety, Dobbe (2022) describes three examples of such concrete feedback mechanisms: audits, accident investigations, and reporting systems. Audits can check whether safety constraints are enforced and whether the system is designed for established practices. In addition to this, accident investigation mechanisms can

provide information for improving the system design, process model, and safety control structure (Dobbe, 2022). Finally, reporting systems allow for reporting hazards before they become accidents which can enable a more safety-aware culture in organizations.

Strategy 7: ***Balancing safety and accountability through a Just Culture.***
System safety practices are fueled by information from the stakeholders involved, their environment, the rationale behind their decisions, and other intangible social factors which should be communicated. A just culture within the organization should be realized for humans to share such information, and insights, or even report hazards. Organizations that run on blame culture and foster an uncomfortable environment risk the change of being unable to properly implement safety practices. In order to do so, organizations should actively foster a safety culture and design mechanisms to maintain it (Dobbe, 2022).

5.4. Main Findings Chapter 5

The goal of this Chapter was to get familiar with systems theory and system safety theory to strengthen the knowledge base. Next to this, this Chapter aims to answer research sub-question four:

SQ4: *What concepts and strategies of system safety theory can aid explainability practices?*

Viewing safety as an emergent system property that must be controlled for shows a new way of approaching elements that are dependent on technical, social, and institutional elements. Safety is both dependent on the reliability of technical components and also influenced by organizational structures and operating procedures. Because explainability is also in need of a socio-technical view similarities to explainability can be established and system safety theory might provide useful concepts and strategies. In order to maintain safety in complex systems Leveson (2003) has created a framework to approach safety based on three main concepts: safety constraints, the hierarchical safety control structure, and process models. Safety constraints can limit behavior in lower levels of the hierarchical control structure and communication channels must be established to communicate the constraints and determine whether the constraint are still being satisfied. Enforcing safety constraints will result in safer systems and installing feedback mechanisms can anticipate environmental and systemic factors. Next to this, the strategies from Dobbe (2022) are relevant for establishing safety within ML based decision support systems and can provide useful insights for approaching explainability. These concepts and strategies will be further incorporated in approaching explainability within the following Chapters.

6. Explainability Approaches Used in Transaction Monitoring

This Chapter is within the third research phase and is situated within the practical environment. The Chapter researches explainability approaches in TM together with the existing challenges and limitations. This Chapter will try to answer research sub-questions five and six, which are as follows:

SQ5: *How are banks currently addressing the needs for explainability within the development and operations of TM systems?*

SQ6: *What challenges arise regarding explainability within the process of TM taking on a socio-technical view?*

6.1. Empirical Study

In order to determine the existing explainability approaches used in local practice at the bank an empirical study has been performed. This study aims to determine the stakeholders that want or need explainability, the existing approaches, and the current limitations and challenges. Semi-structured interviews have been conducted with stakeholders in local practice with a total of 16 participants. Each key role within the TM process has been interviewed, the interviewees are listed in Table 6.1 together with the Interview_ID and link to the according AI system role. Each interview has been transcribed and qualitative data analysis is performed using ATLAS.ti. The codes and sub-categories used for the qualitative analysis are listed in Appendix O.

6.2. Explainability in Transaction Monitoring

Within TM there is a multitude of stakeholders involved concerned with developing, maintaining, operating, auditing, and regulating TM models. The need for good TM systems is clear and has been motivated in Chapter 3 and the operational process and development of TM systems are described in Section 3.4.2 and Section 3.4.3. The actors having some stake in TM is extensive, hence, for the purpose of this research, the most significant stakeholders are selected. These stakeholders are listed in Appendix O Table O.1 and linked to the roles apparent in AI systems defined by Tomsett et al. (2018). Each stakeholder has different motivations and reasons for the need for explainability, the systematic empirical research has tried to identify the most significant motivations, limitations, and risks of the current explainability approaches within TM and will be discussed further within this Section. The scope of responsibilities of the different stakeholders within the TM process is illustrated for clarification in Appendix O Figure O.1.

Interviewee function	TM Stakeholder role(s)	Interview_ID
TM Analyst	Operator, Executor	I1
Consultant Financial Risk Management	Creator-Implementer	I2
Data Science Consultant	Creator-Implementer	I3
Data Scientist	Creator-Implementer	I4
Machine Learning Lead	Creator-Implementer	I5
Data Scientist	Creator-Implementer	I6
Business Developer	Creator-Implementer	I7
Privacy Officer	Examiner	I8
Policy Developer	Creator-Owner	I9
Policy Developer	Creator-Owner	I10
Business Expert	Creator-Owner	I11
Model Risk Management Office	Examiner	I12
Non Financial Risk Manager	Examiner	I13
Operational Risk Manager	Examiner	I14
Product Owner	Not Applicable	I15
Team Lead Innovation	Examiner	I16

Table 6.1.: Interviewees and stakeholder roles in transaction monitoring

6.2.1. Stakeholder Reasons for Explainability

Examiner: External Regulators
The external regulators, for TM mostly DNB and Autoriteit Persoonsgegevens (AP), are concerned with controlling financial entities on their compliance with regulations and national laws. The main regulation on AML, earlier described in Section 3.2, is the Wwft which is enforced by the DNB. The Wwft is more elaborate on explainability and states that banks must understand the workings and rationale of an automated system and must understand the reason for its output of alerts, as regulators might ask for an explanation. The banks are also required to be able to establish the reason why a client has been alerted and must document the evidence of the decision for a SAR filing. Next to this, banks must be able to show which money laundering and terrorism financing typologies are addressed in the TM system. The TM models process enormous amounts of personal data within the training and decision-making. The AP enforces the GDPR and supervises the company's on their data usage, processing, and storage. Within the GDPR the main underpinning for explainability is stated in article 22 saying that data subjects can not be subject to decisions based on automated processing. Within the TM process, there is no automated decision-making because the TM analyst decides whether a client receives a SAR filing. However, the GDPR also focuses on the rights of the decision subject, the client, and states what type of data can be harmful to use and how data should be processed and stored. Because TM systems process enormous amounts of data, the data processing and storage must be explained in order to adhere to regulations. Because the DNB is enforcing the Wwft their reason for explainability is to control the adherence to regulations of financial banks. This is because their responsibility is to oversee and justify the behavior and decisions of banks to the Ministry of Finance of the Netherlands.

Examiner: Internal Regulators
Within the bank, there are two lines of defense that additionally check and validate the design

choices, data usage, and model behavior. The second line of defense consists of the legal, compliance, and model validation team. Legal and compliance are tasked by the bank to internally check whether the TM systems adhere to the regulations such as the GDPR and Wwft. The legal department is focused on interpreting the national regulations of the GDPR and translating these into company guidelines. Elements of the GDPR revolve around personal data storage, usage, and processing. Within the process of developing and bringing a model into production, there are multiple checks performed by the legal department to see whether these guidelines are followed. The compliance department is concerned with providing guidelines for enforcing the Wwft within the bank. Both these departments examine to what extent the TM system is exposed to privacy risks and whether the system adheres to national regulations such as the GDPR and Wwft. The model validation department is also in the second line of defense and is instantiated to validate the design choices of the model developer and the inner workings of the model. Model validation is concerned with checking whether the model's functioning is technically sound. Model validation performs checks on the TM models such as looking into the risks of the model, how the data is processed, what features are used, whether the inner workings follow a structured logic, and many other technology-related aspects. The third line of defense is an autonomous party called audit which does the same activities as legal, compliance, and model validation. The audit department is instantiated as an additional check to determine whether all the processes, models, and operations are compliant and safe. Audits serve as a check for both the risk owner (1st line party) as well as the parties within the second line of defense. The reason for explainability is to control the adherence to regulations and justify the model behavior and decisions at both the operational level and the aspect of data processing.

Risk Owner
The risk owners are responsible for the development and operational activities that are instantiated to mitigate risk scenarios listed in the SIRA. Each risk owner has a team under his management working together in mitigating risks effectively and efficiently. Risk managers are interested in ways to increase the efficiency and effectiveness of both the models themselves and operational activities such as alert handling. Hence, the risk owners have an indirect interest that the analyst can efficiently investigate the alerts produced by the model. Apart from improving the existing models and operations, the risk owner is also interested in discovering new emerging patterns in the data, this can help the overall effort in battling money laundering and terrorism financing. The main reason for risk owners to have explainability is to understand whether the relevant SIRA risks are sufficiently mitigated. Another important reason for the risk owner is to efficiently manage the operations and decrease the costs or time consumed by the TM analysts, this might either be to optimize processes or to increase the efficiency of alert handling by providing increased explainability of the model output. Lastly, risk owners would want to discover new patterns, and to do so explainability might be needed to interpret the driving features or trends in the data.

Data Scientist
Data scientists are tasked by the risk owners to develop models for covering integrity risks stated in the SIRA. Next to this, they are concerned with improving existing TM systems whereas their overarching goal is to cover the risks as effectively and efficiently as possible. The data scientists are the key persons in the TM process because they develop and maintain the TM models giving them a high responsibility for explainability due to their expert knowledge. Data scientists have a multitude of reasons to incorporate explainability, this is because they are responsible for enabling and creating all the explainability tools for the other stakeholders. For the data scientist himself, explainability is mainly empowering control over the model. Data scientists must be able to check whether their model is effectively capturing scenarios for which the

models are initially designed. Also, they need to be able to explain the model behavior, design choices, and model output decisions to risk owners, internal regulators, and external regulators. Explainability can be used to check whether the model is behaving how it is intended or expected to behave. Also, explainability can improve the performance of the model, where for example feature importance methods can improve the selection of features for a model. Another reason might be to discover new emerging patterns within the data. Data scientists can aid in improving the efficiency of operational processes by adding explainability. Examples can be to stop investigating existing patterns or clients that have proved not to be suspicious. Another extremely important reason is to provide the TM analyst with as much guidance as possible on the model output in order for him to make well-informed and understood decisions.

TM Analyst
The TM analyst is assigned to determine whether a client of the bank is suspicious and should receive a SAR filing. The TM analyst receives an alert generated by the TM model, when such an alert is given the TM analyst further investigates the case and performs additional research actions in order to make a decision. The reason for explainability for a TM analyst is to gain guidance for the investigation, it could improve the operational efficiency whenever the model output is accompanied by an explanation stating why the model has generated an alert. Another reason for explainability is that the TM analyst must be able to justify his decision whether the client receives a SAR filing or not.

Client (Business or Natural Person)
A client of the bank, whether this may be a natural person or business, is the subject of the TM process. The TM systems are trained on client data and a client can be alerted. Whenever a client is alerted the TM analyst can perform additional research on the client through existing internal information from the bank or external information from the internet. The TM analyst can reach out to the client in order to gain information on transactional behavior. The main reason for explainability from the client is to be able to control whether the banks are processing and handling their data safely and just, but specifically according to GDPR guidelines.

Data scientist, explainability can:	**TM Analist**, explainability can:	**Risk owner**, explainability can:
- Justify: model behavior and decisions - Control: model objective - Improve: model performance - Improve: workflow of operator	- Justify: decisions - Improve: workflow of operation	- Discover: new patterns - Control: model objective
External regulators, explainability can:	**Internal regulators**, explainability can:	**Clients**, explainability can:
- Justify: model behavior and decisions - Control: adherence to regulations	- Justify: model behavior and decisions - Control: actions by 1st/ 2nd line parties	- Justify: decisions - Control: review GDPR rights

Figure 6.1.: Reasons for explainability of the stakeholders within the transaction monitoring process

6.2.2. Existing Explainability Approaches in Transactions Monitoring

Within TM there are multiple approaches for explainability. The main approaches are model documentation, operational documentation, global model-specific feature importance techniques,

local post-hoc model agnostic feature importance techniques, and information sharing or feedback sessions.

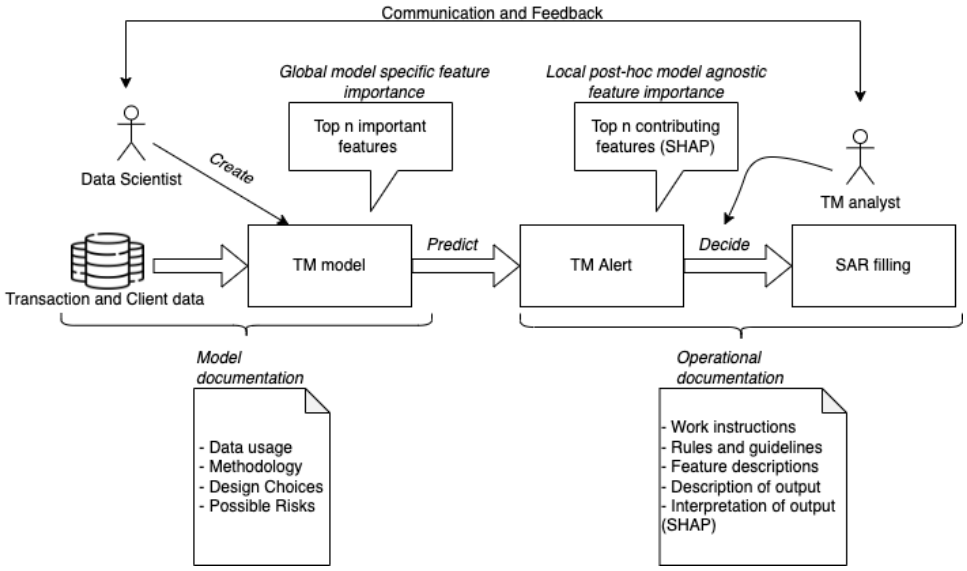


Figure 6.2.: Existing explainability approaches in transaction monitoring systems

Technical Approaches

Global model-specific feature importance techniques

There are global model-specific approaches being used to assess which features are driving a specific model the most, this is used to gain a better understanding of the model and its behavior. This is for example to get an understanding of to what extent features have a relation with money laundering patterns (I3). These methods allow data scientists to compare the most important features over different models and see what the differences are (I2). Global model behavior and feature relevance techniques will provide data scientists with the ability to optimize the model performance, for example with feature selection or the tweaking of features (I2, I6). Apart from optimization, it will allow the data scientists to better communicate the model behavior and high-level workings to other stakeholders either in the model documentation or in the communication sessions.

Local post-hoc model agnostic feature importance techniques

Data scientists mostly make use of post-hoc model agnostic methods to provide TM analysts with additional information on the model output that can aid them in deciding to file a SAR or not. The most widely accepted and used method is SHAP which scores the features that have the most contribution towards a certain model prediction, in this case, the prediction to alert a customer. SHAP is considered effective in determining how features are driving the score for the alerted clients. SHAP selects the top n features that contributed the most in alerting a client. Apart from existing approaches, data scientists try to develop local post-hoc methods tailored for specific TM models. The models are also focused on feature relevance techniques.

Social approaches

Within organizations, there is an unmeasurable amount of communication that can either be formalized or unformalized. The TM landscape can be represented as a hierarchical structure

going from the higher-level stakeholders which are the external and internal regulators all the way down to the TM analysts. Often communication takes place, or is most important, between two neighboring levels. This is because the higher level can impose constraints on the activity of the lower level. Figure 5.1 shows an example of the relation between two levels that are represented by the data scientist and TM analyst. Now, the most important social approaches to the explainability of TM models will be discussed.

Model Documentation
Model documentation serves as the main communication to explain the applied algorithms, global behavior, and inner workings of a model to any stakeholder or to other data scientists. Model documentation also describes what data will be used and how. Next to this, it includes potential privacy risks and explainability risks. Originally documentation is created for your future self, in case you need to revisit the design choices and circumstances at the time of the design. However, within cooperations, documentation is also extremely important for others. Documentation will allow other stakeholders to assess the model without requiring technical expertise, it will allow them to test their own requirements upon the design choices, and allow them to compare different candidate models across the axes of ethical, inclusive, and other non-traditional evaluation metrics. Especially within the case of a high decision-making area such as TM, it empowers policymakers and regulators on asking questions about a model and knowing benchmarks around the suitability of a model in a given set (Mitchell et al., 2019). There are established guidelines that foster good documentation practices such as the *Model Cards for Model Reporting* from Mitchell et al. (2019).

Operational Documentation
The operational documentation serves as a document that explains the output of the model and how this can be interpreted. The operational documentation is specifically meant for the TM analyst and can aid the analyst in understanding the meaning of features, explainability methods, or other kinds of information that will help him in a better decision-making process. Besides providing guidance for an improved decision-making process, the operational documentation also includes the work instructions and procedures the analyst should follow which represent the company and regulatory guidelines. The operational documentation is similar to the *reference channel* from Figure 5.1.

Communication and feedback
Two examples of formalized communication channels between layers are already given with the model documentation and operational documentation. However, there are also communicative means that are less formalized but not less important. Such means can be meetings, feedback sessions, or other conversations all serving the purpose of providing insight and understanding of elements of the TM process. Discussions can improve workflows, and efficiency, and align requirements. Also, they are the most human way of explaining and are a great means to stimulate dialogue. These forms of explainability approaches serve as derivatives from the earlier described approaches, they often function as additional or explanatory information on top of the existing approaches. Within banks, there are multiple forms of communication that fall within explainability approaches. The three main tools for communication are meetings, feedback sessions, and checklist forms.

Meetings are the least formalized and can consist of conversations between the regulator and the risk owner to explain the techniques used in the existing approaches. Such conversations can help to determine whether this fits the regulatory guidelines. Next are the feedback sessions, these are for example discussions where a TM analyst asks for additional information on the model behavior or certain features. Note that these questions are initiated by the analyst

himself and can be seen as the *measuring channel* pictured in Figure 5.1 (I2). And then there is the most formalized way of communicating and this is through the checklists, these are often part of internal regulatory guidelines and are installed to let the data scientist explain the model, its capabilities, the risks, and potential impact and additionally determine whether these fit with the GDPR, Wwft and organizational guidelines. The checklist is communicated with the internal regulators such as the 2nd and 3rd line of defense. Finally, there is a form of communication from the bank towards its clients explaining on a high-level how their data is used and what kind of methodologies or techniques are applied.

6.3. Environmental, Organizational, and Behavior Shaping Factors within the TM Landscape

Human behavior is greatly influenced by the context and environment in which the human is working. These behavior-shaping mechanisms should be taken into account when considering real-life situations (Leveson, 2011). Such behavior-shaping mechanisms are influencing the TM landscape, banks, their practices, and the behavior of employees. There are a few factors that influence explainability practices which will be described below.

- **Regulatory pressure:** there has been an increased pressure on banks to ensure that their TM systems are fully compliant with the requirements from the Wwft. This pressure is coming from the regulator and has resulted in the need for developing the systems as quickly as possible to adhere to regulations (I8). The pressure is translated into fines and the possibility for management to be sued (I8).

- **Key person risk:** this is the loss of tacit expert knowledge. There are multiple factors that constitute key person risks, especially within the financial sector and banks there is a high turnover of employees which is becoming a big problem (I5). Factors influencing this includes the banks competing for each other's employees, a general employee shortage, and the repetitiveness of the work of the operator (I8, I5, I14, I15).

- **Capacity problem:** there is more work than can be processed. Due to factors such as regulatory pressure and key person risk, there is a capacity problem (I3). The regulatory pressure is forcing banks to comply with regulations resulting in catching as much suspicious activity as possible, however, this results in a capacity problem when taking into account the high employee turnover (I3).

- **Novelty:** explainability is a topic that has been given a lot of attention recently, but there is not yet an established view on the implementation and adoption within the industry (I2). This results in the lack of both knowledge and awareness of explainability and the lack of available methods. Next to this, the regulators are still figuring out how to regulate the topic of explainability and do not have yet clear guidelines and requirements on how to incorporate this within organizations (I3, I5, I8).

- **Misconception of AI capabilities:** AI is viewed as a crystal ball that can perform any task (I10). In order to set realistic requirements the limitations of techniques should be known. AI is not the solution for every problem, often people have the tendency to incorporate additional or new techniques which are not tested properly.

- **Profit Motive:** organizations have a profit motive and have the priority to generate as much revenue as possible. This can create friction with ethical considerations or limit innovative practices or research.
- **Increased Complexity:** due to the unconstrained technical ability of today's software-based systems, there is a tendency to incorporate more features, elements, or products within the system. When elements are not removed or made more efficient, this will inevitably lead to complex systems over time.

6.4. Concerns and Limitations About and Across Existing Explainability Approaches

Within the interviews all five explainability approaches have been discussed, this resulted in multiple concerns and limitations of the methods themselves but more importantly on the usage of methods. There are multiple factors that influence an explainability method, or its usage, and are often technical, social, or organizational. A few of the main driving factors leading to risks have already been described in Section 6.3. Often a combination of factors can lead to risks, these can be either environmental, behavior-shaping, or limitations of the method itself. These risks are imposed by the method and whenever unanticipated lead to hazards that can further affect the entire TM system. The hazards are specified as direct hazards, which represent scenarios that directly lead to a loss. Or unanticipated risks can lead to other factors, which do not directly cause loss but can foster risks or new direct hazards over time. These factors that arise from the unanticipated risks often find their origin in system safety theory, an example of such a factor is the misalignment of mental models. A structured overview is given in Figure 6.3. Each approach will be discussed starting by describing the direct risks, followed by the implications and influence of the risks on the entire TM system by discussing the hazards and arising undesirable factors, and finally, the factors influencing or creating these risks are described. Each section will also have a discussion on possible improvements for the explainability methods or usage of the methods.

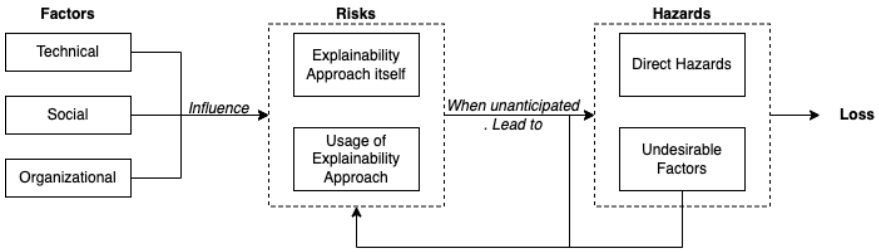


Figure 6.3.: Elements included in the concerns and limitations of explainability approaches

6.4.1. Model Documentation

Risks and Hazards
There are risks involved with model documentation practices observed in local practice. Model

documentation can be viewed as a meta-explanation instrument that needs to represent the actual model behavior and includes the design choices and the rationale behind them. Therefore, the main risks of model documentation itself are when there is a misrepresentation of the actual model or choices, there are missing elements, the representation is unclear, or when the documentation is not updated accurately. These risks reside in the fact that there are no general or management practices established to generate uniform documentation, validate the understanding, and control and update for changes.

The factors that can arise from inadequate documentation or documentation management that do not directly cause hazards, but can enforce them, are the misalignment of mental models and requirement errors. Whenever documentation is the only source of information for stakeholders, a misconception of the actual model can be established which can lead to the specification of flawed follow-up requirements or design choices. Also, their risks or the capability of the model can not correctly be accounted for when there is a misrepresentation of the actual model behavior. Whenever model documentation is not revisited and updated this can lead to entirely unexplainable design choices, in such cases a model in production is not entirely understood which brings the entire TM system to a higher state of risk. Such direct hazards can be dangerous and eventually, as observed in practice, can lead to the discarding of entire models due to the fact that the model documentation was unclear resulting in the inability to explain the model predictions or feature relations (I8). Although cumbersome, often in such cases developing a new model is better than trying to interpret the unclear documented one. Good documentation practices are crucial to avoid these risks (I13).

Factors
There are a lot of factors influencing the risks of unclear, missing, un-updated, or low-quality documentation. A selection of factors that have been observed in local practice will be discussed, the most important factors are the regulatory pressure, fast-changing environment, and high turnover of employees. The TM landscape within banks has been set up rapidly due to regulatory pressure from the DNB which can have consequences on good practices of documentation (I8). Within such a dynamic environment it can be hard to perform deliberate documentation, however, documentation in such scenarios is even more important. Another factor influencing documentation practices is the competitive landscape of banks where there is a high turnover of employees resulting in key person risk (I5). Model documentation is created by the data scientists themselves, however, when the model developers are externally hired parties or when the turnover of employees is high it can lead to insufficient documentation and the inability to update or revisit the rationale behind design choices. This can happen when there are no existing practices installed to revisit or update the model documentation. Another example is when models need to be retrained in order to anticipate data shifts, the model can pick up new suspicious patterns from clients. When such global behavior change in the model is not incorporated in the model documentation this leads to wrong assumptions and is an example of a system migrating to a higher state of risk over time. Therefore, it is extremely important to revisit and update the model documentation over time.

Possible Improvement
To determine how good documentation practices should be established the actual goal of documentation should be revisited. The goal of good documentation is to align mental models, everybody that uses the documentation should have the same clear understanding independent from prior knowledge of the model. Therefore, it is important to establish a certain entry level of explainability for the model documentation based on the intended audience. The audience and goal of explanations should be determined beforehand and model documentation can

be tailored to these properties, this can imply different versions or elements for different stakeholder model documentation. Another practice could be to instantiate uniform elements as part of the model documentation so that this can become more familiar and easily interpreted over time, also this will help model developers in being able to continue to work on models from others (I7, I8). Model documentation should be a continuous and living document that is adequately representing the actual state of the model behavior and perfectly reflects the design choices. Apart from structured model documentation practices, there is a need for establishing a management structure that allows for controlling and validating whether the documentation is still representative of the actual situation.

6.4.2. Operational Documentation

Risks and Hazards
The risks with operational documentation are similar to that of model documentation. Because operational documentation is a meta-explanation of certain model elements and additionally consists of work instructions, there are the risks of misrepresenting the actual scenario, missing important elements, unclear or too complex documentation, and updated documentation. In practice, it is quite often the case that the normative work instructions and rules deviate from the established practice. This is because operations are much more receptive and able to adapt to changes in the actual system and its environment.

The undesirable factors that can come from these risks are that when there is already an unanticipated gap between the normative procedure and established procedure, there exists a misalignment of mental models which will get worse over time. There are also multiple hazards that come from insufficient operational documentation practices. When the operational documentation is wrong or unclear it can hamper the TM analyst in making informed and adequate decisions. Inadequate decision-making of the analyst eventually leads to a loss, which is represented by a false positive. There is also the hazard that operational documentation while being correct becomes too large or complex to be usable for the operator (I14). When operational documentation is not taking into account the environment of the operator and behavior-shaping mechanisms such as the capacity problem and high work pressure this could result in the operator's establishing their own procedures. Apart from the fact that this can lead to operator error, this can also force the operator to make decisions based on his own mental model based on previous experiences which can reinforce existing biases (I5). An example is a scenario where the operator must investigate all transactions that the model provided with an alert, this is a trade-off between selectivity and relevance (I2). However, when the normative procedures do not match reality, showing that the activities are not effective can lead to the operator deviating from the rules and being guided by his own experience.

Factors
There are a lot of factors influencing extensive, misrepresented, or wrong operational documentation. One factor is that the operators are not involved in designing the operational documentation and the other way around is that established practice is not directly communicated to the developers. Another factor is that over time the TM system is becoming more complex due to the lack of technical constraints, this will result in exhaustive and more complex operational documentation. Such documentation could be misrepresenting the actual scenario and be unrealistic to take into account within day-to-day operations. Other factors that could influence the deviation between established practice and operational documentation are factors such as time pressure and cost efficiency. When TM analysts are under time pressure operators will try to

become more efficient and productive to deal with the time pressure. This will result in a growing deviation from established practice and the normative work instructions and rules because the idealized procedures often do not include such factors. Next to this, the operator his ability to interpret and efficiently work with the operational documentation depends on personal and task-specific factors. Research from Wagner (2019) states that, among other factors, the amount of time available for the task, the degree of qualification of the operators, the degree of liability of the operator, and the ability to change the decision all influence the behavior of the operator. Next to this, when operational documentation does not provide the correct level of support for the task designed for the operator and provide access to all relevant information the decision-making system can be defined as quasi-automation (Wagner, 2019). All the factors mentioned above should be taken into account when developing operational documentation in order to align the operational documentation with the characteristics of the TM analysts and resemble the actual working environment.

Possible Improvement
In order to develop operational documentation that is accurate and representational for the environment, the mental models of the TM analyst and developer should be aligned. This can be done by enhancing the communication and feedback channels from the operators to the developers to give them insight into daily practice and discuss edge cases. Such measures can calibrate the knowledge of both and can close the gap between the normative procedure and the effective procedure. Next to this, environmental factors such as working under time and productivity pressure should be taken into account. Taking lessons from the experience of daily interaction with the system makes the TM analyst so valuable.

6.4.3. Global model-specific feature importance

Risks and Hazards
The risk with the global model-specific feature importance techniques is that they can often fall short in entirely representing the model behavior. This is because the global model-specific methods are only showing which features drive the model the most based on the existing set of training data. However, models can shift in behavior due to the dynamic environment and changes in data. Because the methods are limited in representing the model behavior it is hard for developers to control the objective of the model. A lot of concerns are being raised by practitioners on this topic. Models are being developed to mitigate certain scenario's from the SIRA, however, to what extent the models are actually capturing these events is hard to quantify (I3, I5, I6, I9, I11). Such global model-specific behavior should be known and current explainability practices are unable to provide this information. Now the models are only controlled for catching SIRA events, but it is not clear what events are being captured.

The unidentified global behavior or misrepresented model behavior can lead to wrong assumptions that result in an inaccurate mental model. This again can lead to flawed requirement specifications for the operational documentation for example. Being unable to control the designed model objective raises questions of usefulness, especially within an environment that optimizes for costs such as within financial banks (I6). This results in the challenge to quantify the effectiveness of the designed models (I6). Next to this, there is a more ethical underlying question of whether models should be used when we can not exactly validate which scenarios they are actually mitigating. Questions have been raised by practitioners about whether it makes sense to use such complex models when it is not clear what is being mitigated, should these models be discarded (I9)?

Factors
Global feature importance methods can not handle feature correlations well. An example is whenever feature A and feature B are correlated and A is linked to suspicious behavior but B not, it is still possible that B (through its correlation with A) takes on higher feature importance, and the wrong conclusion can be drawn. Also, the methods are limited to one model and require technical knowledge of the existing system and explainability technique applied.

Possible Improvement
Similarities can be observed with the alignment of established practice and effective practice for the work of operators. Whenever a model is designed for mitigating certain SIRA risks this should be controlled and validated. By implementing control and validation structures the output and mitigating measures of the model can be explained (I9). In the case that global model-specific explainability techniques fail to do so additional measures should be designed. When the model is in production feedback channels could be leveraged using the tacit knowledge and experience of the TM analyst to validate whether SIRA scenarios are caught in the output of the model (I5, I9). The most important takeaway is that explainability must be enforced by controlling whether the information the model is designed to monitor is coming back in other places within the process (I9). However, such techniques, that try to catch global model behavior, are currently not in place within the TM landscape. This has implications for the entire adoption of ML models because they can not entirely be relied upon, and has resulted in the fact that the rule-based models and ML models are still being run simultaneously.

6.4.4. Local post-hoc model agnostic feature importance

Risks and Hazards
There are a few risks coming along with local post-hoc model agnostic feature importance methods such as SHAP being used in TM. Concerns have been raised that SHAP is lacking and that the existing methods are incomplete. Practitioners found that SHAP itself can be seen as a black box because the game theory and analogy behind it are not entirely clear (I3). The risk of using such local explainability methods is that they can result in a wrong image of simplicity or provide a false sense of trust (I5). When providing the analyst with the top 3 contributing features can misguide the analyst which is cost-ineffective. Currently, practices have decided upon showing n features with the highest SHAP values, however, setting a fixed number for this may provide too much or too little information. There is a trade-off between selectivity and relevance. Also, it can influence the mental model of the analyst who can start over-relying on the explainability tool or on features or clients that are provided on multiple occasions (I7). Also, when handing the analyst a wrong or incomplete explainability this can result in the filing of an alert where there could have been fraudulent behavior, resulting in inadequate decision-making.

Factors
SHAP still has technical limitations such as the inability to handle correlations between features entirely (I2, I5). Other factors leading to wrong assumptions, over-reliance, or misinterpretation of SHAP is that these technical limitations are often not known. When limitations are known and acknowledged, it is unlikely to entirely rely upon such an additional explainability tool. Next to this, there is a lot of pressure on analysts to handle alerts because of capacity problems, which may lead to a forced over-reliance on explainability tools (I3).

Possible Improvement
Methods should be tested before being applied in practice, examples could be to test methods

on simpler models to see whether they produce the expected output. However, it should be noted that methods such as SHAP become less reliable when the complexity of the features and feature interactions increases. Therefore, such validation should be developed with care. Also, the limitation and interpretation of local explainability methods should be clearly defined and communicated to avoid over-reliance or misinterpretation. When discussing the limitation of current local explainability approaches there are other overarching concerns being raised, as one practitioner stated: *"Currently, there are no available tools or controls to see to what extent the TM analyst is relying on the local explainability methods, when they do overly on it and handle alerts differently than the model is intending, it begs the question whether the model is built correctly? Such an issue must be approached from a systems perspective (I5)"*. When local explainability methods can not work well for the highly complex models and features themselves become harder to understand the explainability becomes harder to achieve but also much more important (I2). Concerns have been expressed on the complexity of the models in general and whether the limits are being pushed (too far) (I2, I3, I5).

6.4.5. Communication and Feedback

Risks and Hazards
The risks of the communication channels, such as checklists, are that they might provide a sense of false security. It is important for the person conducting these checklists to be able to really understand the model, its capabilities, and risks (I11). Apart from the checklist serving as a reminder for the data scientist, it is important to actually oversee and validate the design decisions. Whenever a checklist is not supervised it often just serves as paperwork without real meaning. For the more unformalized communication means such as feedback sessions and meetings, there is the risk of employees that do not want to set up a meeting because the other parties might be busy or because it can feel like the questions will be a burden. Also, within the sessions themselves, it might be difficult to really understand or probe the reasoning of the data scientist because of knowledge asymmetry (I11). A lack of communication can result in unclear guidelines, currently, the guidelines on AML and explainability from the DNB are described vaguely which results in banks interpreting this on their own (I8, I5). This can result in the risk of doing too much or too little, or even the wrong thing.

Factors
The main factor that influences limited initiations of feedback sessions or being willing to admit the knowledge gap and probe the reasoning of the data scientist, is the company culture. Another factor that might enhance such fear is the growing misalignment of mental models. Whenever the TM analysts feel that for example within a feedback session, they do not actually understand what is happening they could tend to leave it for what it is. This factor is similar to the checklist or regulatory discussions, whenever the other parties have a limited understanding of the techniques being used or understand this then the communication is just a form of self-reflection and not a discussion (I11). Next to this, the feedback and interaction between data scientists and operators might be hampered due to capacity constraints (I3). Especially, the senior analysts are the most valuable and can contribute to knowledge sharing but these are the ones with the least time available (I3).

Possible Improvement
One of the most important things to establish within an organization is a good and safe company culture. There should be an open culture where employees are not afraid to ask anything or have the feeling that they are hampering someone's productivity. Also, it is important that

there exists a culture where nobody feels ashamed to admit some knowledge differences in order to level it and do something about it. Apart from checklists serving as a check, they should actually be probed and questioned in order for it to really work. To achieve this the mental models should be aligned so that a shared understanding will be created. This will allow different parties to communicate requirements and know what is actually happening within the work of others on the topic of certain techniques (I17). Also, a standard format can help to align thinking and structure the communication so that both parties know beforehand what will be discussed (I8). Feedback is critical for good design practices and the safe operations of the analysts. A basic principle of system theory is that no control system will perform better than its measuring channel (Leveson and Weiss, 2009).

6.5. A Novel Approach to Explainability in Transaction Monitoring Systems

There are multiple risks that can arise when technical explainability approaches, meta-explainability approaches, and social explainability efforts fail or are not well executed. The empirical data has shown that practitioners within TM, although having considerable attention to this topic, do not always succeed in designing the right explainability approaches that can fulfill the needs and desires of the stakeholders. In addition to this, there are both technical factors, environmental, and social factors play that heavily influence the environment of stakeholders which can have an effect on the approaches they develop or use. Next to this, the TM system itself and the environment changes due to the dynamic complexity of the system (Leveson and Weiss, 2009). As explainability is more than needed in the complex and dynamic environment of TM this calls for an approach that takes into account the technical limitation, social factors, and both organizational and systematic factors. Currently, such an approach is missing. There are multiple efforts that try to realize explainability, but these efforts do not explicitly take into account the audience, stakeholder desires, and external factors.

The existing efforts approach explainability from a technocentric perspective, but explainability has socio-technical properties and is applied in a highly complex system. Therefore, explainability should be addressed by using a socio-technical approach using systems theory (I5, I7). To ensure that the entire TM system is explainable, the social, technical, and institutional elements should be taken into account during development. The empirical research has shown that apart from the limited technical abilities of existing XAI, explainability is mostly approached with engineering development efforts. And, similarly to safety in software systems, the important elements for ensuring explainability are met through setting clear requirements, complete up-to-date documentation, and verification and validation (Leveson and Weiss, 2009).

This Section suggests the need for a user-centered operationalization of explainability and socio-technical control and validation. Currently, ad hoc and loose efforts do provide value and fulfill certain stakeholder needs. However, there is no clear approach on how to tackle explainability system-wide and measure the fulfillment of requirements of whether problems are actually solved. Operationalization will allow for validation and control and will in the end result in an improved, measurable, and more structured approach to explainability.

6.5.1. A Need for User-centered Operationalization

Explainability is addressed by multiple approaches consisting of: model documentation, operational documentation, global model-specific feature importance methods, local post-hoc model agnostic feature importance methods, and means of communication and feedback. The explainability practices and motivations are fueled by the reasons of different stakeholders for explainability pictured in Figure 6.1. The main reason for the existing explainability approaches is because of the need to adhere to Wwft, and by doing so justify the inner workings and decisions of the TM model. Although there are many more explicit motivations for explainability from different stakeholders, these are not taken into account during the development and deployment of current explainability methods. Local practitioners clearly mention the need for incorporating the audience of explanations together with their reasons and desires to establish an optimized design.

Currently, explainability is addressed ad hoc, without taking the audience or their reasons into account during design and development, and through loose efforts that are not coherent, structured, and controlled over time. Explainability is a social process and approaches are influenced by social, technical, and environmental factors. Observed in practice, the usage and effectiveness of explainability approaches are highly affected by factors on the social, technical, and environmental levels. Current efforts do not structurally take these factors into account. Next to this, practitioners suggest that in certain scenarios explainability should not be necessary when the design choices of the initial model concepts can also be changed in order to fulfill the stakeholder needs. This shows that explainability is not addressed in an inadequate manner, Adhoc and through loose efforts and instruments that do not necessarily provide an effective fulfillment of what is actually needed. Explanations are becoming meaningless when they are not considering the audience, usage, and reasons. Defined by Miller (2019), explanations are an answer to a why-question, and to satisfy this both the question and objective must be known.

This highlights an existing gap within empirical research and shows the need for a structured approach to explainability in order to fulfill the explainability needs and reasons of the involved stakeholders while taking into account the technical, social, environmental, and other emerging factors. In order to establish this, this research proposed the need for a user-centered operationalization of explainability.

6.5.2. A Need for Socio-technical Control

When approaching explainability within the TM system environment, there are multiple interactions between stakeholder groups that each have different reasons for explainability. Also, there are multiple explainability approaches that try to fulfill those needs. The complex environment of TM, using software and intelligent ML models, can result in undesired scenarios that emerge from interactions between components, system requirement and design error, misalignment of mental models, and indirect interactions and systematic factors leading to factors that are limiting current explainability approach (Young and Leveson, 2014). Explainability can only be determined and controlled in the context of the whole system, therefore, this research proposes that explainability is an emergent property of systems. Explainability can not be approached by only considering the single component it covers, but explainability arises between different levels of the system and through interactions.

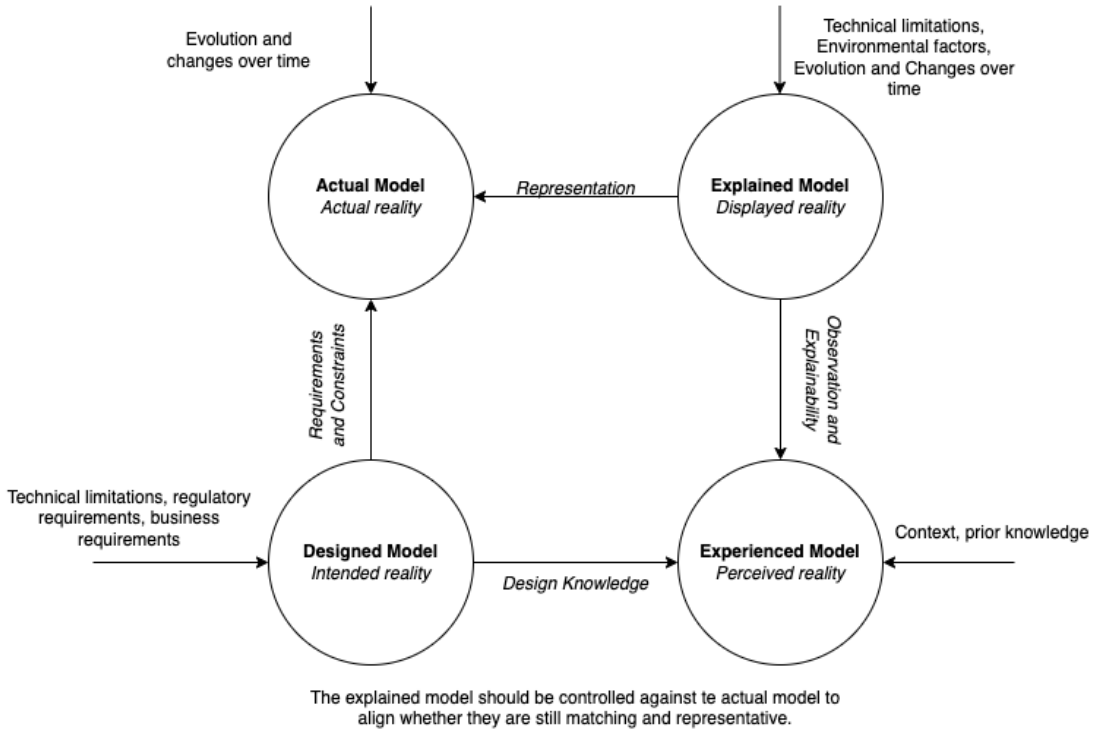


Figure 6.4.: Four different types of the same models

Most of the existing explainability approaches have the risks of misrepresentation. Because explanations are meta-instruments that describe an underlying phenomenon, the explanation must change accordingly whenever the actual system, model, or elements change. Local practitioners have described that there exist no structured means that validate or control whether the explainability approach is still representative of the underlying model or element. When there is an unanticipated misrepresentation, for example with the model documentation, this can result in misaligned or inaccurate mental models, flawed requirements, or inadequate decision-making which eventually can result in losses. On the other hand, explainability is used as a control mechanism checking whether the model behavior represents the designed behavior for example. Taking a look at Figure 6.4, it can be seen that the explainability approaches create the perceived model. In such cases, when the explainability methods themselves have unanticipated limitations the control is malfunctioning. Both cases ask for instruments that validate and control whether the explainability approach itself or the explanation is still representative of the underlying explained element, or model. Therefore, this research poses explainability as a socio-technical control problem. An explainability approach itself is subject to environmental, technical, and social factors and must be controlled for whether it still accurately represents the actual model.

6.6. Main Findings Chapter 6

This Chapter has been set out to answer research sub-question five and six and the results will be discussed below.

SQ5: *How are banks currently addressing the needs for explainability within the development and operations of TM systems?*

The empirical study has shown that there is a multitude of reasons for explainability covering reasons to justify, control, improve, and discover. The existing approaches to explainability are model documentation, operational documentation, global model-specific feature importance methods, local model-agnostic feature importance methods, and communication provided by feedback sessions and checklists. The main method used for local model-agnostic feature importance is SHAP.

SQ6: *What challenges arise regarding explainability within the process of TM taking on a socio-technical view?*

Within operations and development of TM the explainability methods give rise to significant risks and concerns. The empirical study has shown issues regarding the misalignment of mental models due to incomplete or unupdated model documentation which can result in flawed requirements and flawed design choices. Often this occurs due to asynchronous evolution for example when the model is retrained and the newly detected patterns are not included in the documentation. Next to this, the operational documentation can become too complex due to creeping featurism or because the mental models of developers and operators and not aligned. In such scenarios, the development of the operational documentation is guided by the intuition of the designer and causes a gap between the normative procedures and the established procedures. For SHAP there is the risk of over-reliance and a lack of understanding by the operators into the limitations of the method. This can result in simplifying the result and can lead to inadequate decision-making. Next to this, communication approaches such as checklists are often too technical and hard to understand for stakeholders, this can result in the inability to probe the rationale behind design choices for non-technical stakeholders. These are examples of challenges regarding the explainability approaches within TM. The main causes that drive these risks and challenges are that the intended users of the explainability approaches are not taken into account within the development and that the approaches are not controlled and validated over time. Next to this, the main limitation of existing explainability approaches is that they are tackled through loose efforts that merely focus on technocentric solutions and do not take environmental, social, and organizational factors into account. This research proposes to solve this by situating explainability as an emergent system property that must be controlled for. This must be done by developing a user-centered operationalization of explainability while instantiating socio-technical control.

7. A Method for Operationalizing Explainability

This Chapter is situated in the fourth research phase and combines the insights from the practical environment together with the empirical study and the knowledge base including the systematic literature review. The Chapter will start by introducing the novel view on explainability as a socio-technical control problem and will try to answer research sub-question seven:

SQ7: *What method can guide the operationalization of explainability within ML based decision support systems, taking on a socio-technical view?*

7.1. Explainability as a Socio-technical Control Problem

Chapter 6 proposes a new view on explainability stating that it is an emergent system property that must be addressed as a socio-technical control problem. The empirical study has shown that the definition of explainability is the fulfillment of explainability constraints that come from the stakeholder requirements (i.e. reasons) while anticipating and controlling for additional factors that may influence the environment. Next to this, the empirical study shows that explainability is currently addressed in an inadequate manner, the approaches are technocentric, ad hoc, and through loose efforts and instruments that do not necessarily provide an effective implementation. Whereas an effective implementation is defined as the fulfillment of the stakeholder's requirements.

Therefore, this research proposes to approach explainability by i) establishing a user-centered operationalization providing a structured approach to develop practical system constraints with ii) a socio-technical control structure ensuring these constraints can be satisfied over time. Operationalizing explainability will provide a structured approach for establishing explainability practices that are resilient in dynamic complex systems taking the users and their processes into account. The operationalization will provide explainability system requirements that can be incorporated within organizations covering institutional, technical, and social elements that will be combined to instantiate and control explainability. Developing an operationalization and control structure ensures that the explainability of a system can effectively be managed, validated, and controlled.

7.2. Operationalizing Explainability

The empirical study has shown existing gaps in meeting the user's needs and reasons for explainability and points out that current explainability approaches rather focus on algorithmic-centered practices which do not take social and environmental elements into account. A user-

centered approach will take the audience for which explanations are intended into account rather than merely the intuition of the designers that develop the explainability approaches.

The method for developing explainability constraints can be dissected into three categories, it tries to understand (1) the *who* referring to the processes and stakeholders, (2) the *why* entailing the needs and reasons, and (3) the *what* which involves the explainability means, criteria, and environmental factors. These elements will provide an actionable trade-off that, while providing limitations, will result in explainability system constraints. The method of operationalizing explainability into system constraints is presented in Figure 7.1.

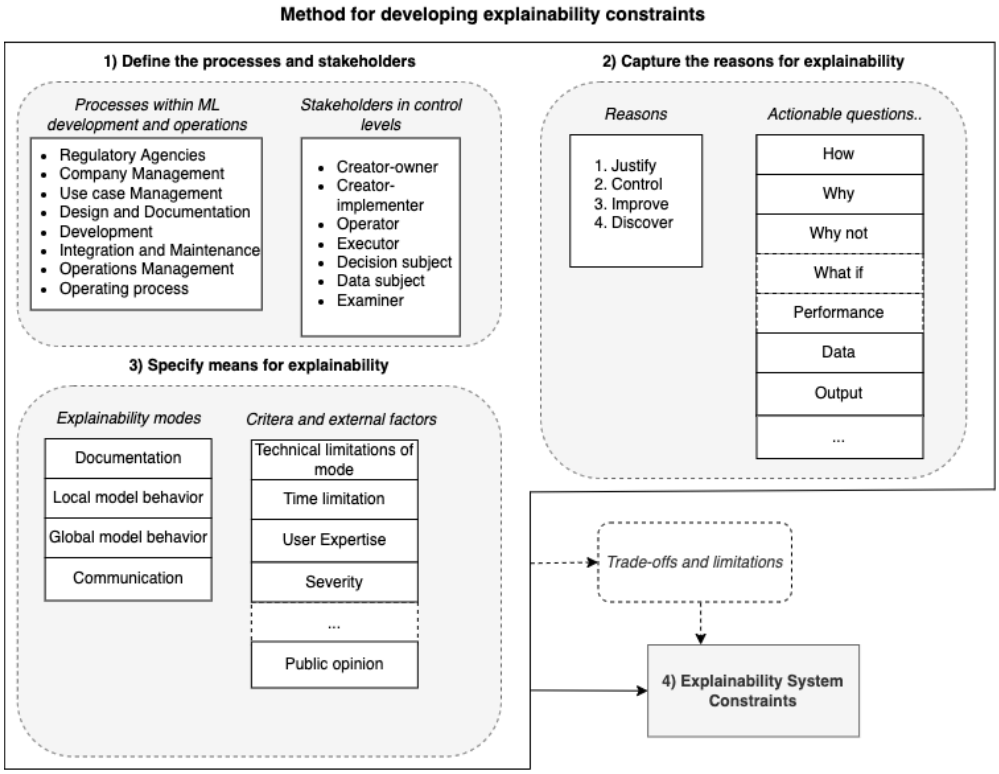


Figure 7.1.: Establishing explainability constraints from user requirements

The *who*, *why*, and *what* are incorporated within the following three elements of the method.

1. **Define the processes and stakeholders**
2. **Capture the reasons for explainability**
3. **Specify the means for explainability**
4. **Explainability System Constraints**

These three steps all have sub-elements and will be further elaborated within the following Subsections.

7.2.1. Define the processes and stakeholders

The *who* refers to the stakeholders that are active within different processes of the ML system. This can either be within ML development or operations but also within project management and regulations. These processes can be found in the hierarchical control structure, an example of such structure is provided earlier in Figure 5.2. Often explainability is needed within processes and between processes, or when talking about the hierarchical control structure between control levels. Taking the processes as the main focus, stakeholders can be identified that operate within these processes. Together these make the source for any requirement and reason for explainability. It is important to focus both on the processes and the stakeholders, this is because stakeholders are often active within multiple processes so a clear delineation will provide clarity on where the reason for explainability lies.

As addressed multiple times, explainability is a social process that is highly dependent on the intended process or audience holding the reason for explainability. So, the first step is to identify the processes in which explainability is needed or where it can empower stakeholders in their actions. To identify the processes the hierarchical control structure of the organization must be mapped to determine where the exact reason for explainability might be present and which stakeholders are interacting. Within the method, pictured in Figure 7.1, example ML system processes are provided.

Now that the processes are defined the stakeholders within the control levels must be determined, this is done so that the reasons for explainability can be extracted. For this example, the AI systems stakeholders described in Subsection 4.4.1 are listed. These stakeholders all have different responsibilities and actions they want to perform, within the processes, which explainability can empower them to do.

7.2.2. Capture the reasons for explainability

Each process and involved stakeholder group has different reasons for explainability that they need to fulfill their responsibilities and empower their own actions, these reasons can be categorized and are: to justify, control, improve, and discover. Within these categories, the specific scope can vary focusing on certain parts of the ML model, data, or process.

7.2.3. Dissect Stakeholder Reasons into Actionable Questions

To extract and concretize what is actually driving the reason for explainability it must be turned into actionable questions. Developing actionable questions will help stakeholders to materialize what they want and need to know, this can be fruitful especially because stakeholders often do not exactly know what they want to know. Next to this, it will provide divide the reasons into elements that will be easier to act upon and can result in increased expectation management. Taking inspiration from the XAI question bank from Liao and Varshney (2021), some example questions are listed in Figure 7.1. Examples of how some of these questions might be manifested are provided below:

- **How:** asking about the global model behavior or the general logic or processes the ML algorithm is following.
- **Why:** asking about the local prediction and the rationale behind it.

- **Why not:** asking why a local prediction is a certain way and not another.
- **What if:** asking how the input relates to the local prediction.
- **Performance:** asking about the performance of the model.
- **Data:** asking about the data usage, or how the model is trained (e.g. protected attributes).
- **Output:** asking how the output can be interpreted or what the procedures are to handle this.

7.2.4. Specify the means for explainability

Specifying the means for explainability will help to determine what is possible to answer the actionable questions and will take the criteria for explainability into account. These criteria often come forth out of environmental or organizational factors. The explainability modes, described in Figure 7.1, are coming from explainability practices observed in the practical environment of TM. These modes are local model behavior techniques, global model behavior techniques, documentation techniques, and communication. It is important to note that there are scenarios where the explainability mode is not within one of these four classes.

When taking into account the mode of explainability, there are certain criteria that come along such as technical limitations, the quality of produced explanations, time available to develop and maintain the approach, or how resilient the approaches are to changes over time. Whenever choosing an explainability approach, these must be tested against elements such as the properties of explanations methods described in Subsection 4.2.5 and properties of good explanations described in Subsection 4.2.4. Technical abilities and limitations must be known in order to select the right approaches and to design processes around them in order to make them reliable.

Next to this, other criteria must be taken into account. Discussed in Subsection 4.2.3, there are latent dimensions of explainability that need to be taken into account which must be approached as criteria. Insight from the empirical study has shown that less measurable elements can greatly influence explainability needs. These criteria eventually influence the practical implementation of explainability and must be incorporated within the design of explainability constraints. The criteria are part of the assumptions used to derive the eventual design features and system requirements, therefore, it is important to clearly document these. If the criteria, or assumptions, change over time or the system changes and the criteria are no longer true, then the constraints need to be revisited. A few examples of such criteria are provided below.

- **Time limitation:** is about the time the user needs or is allowed to have to gain an understanding of the explanation. Within a decision-making process, this might be short or when auditing a model based on the model documentation this can be longer.
- **User expertise:** users of the model might have different backgrounds and knowledge of the existing system or techniques used, this might influence their understanding and also the appropriate detail of the explanation.
- **Severity:** what are the consequences whenever the explanations are wrong or incomplete, this should be determined in order to prioritize the eventual constraint.

Finally, there are other external factors that can not be controlled but do play a role in the design and need for explainability. Although, these factors are hard to capture an example is the public opinion. Public opinion changes over time and is hard to measure, however, when the public debate pays extra attention to XAI or the need for more explainability within ML based decision support systems this will put pressure on the explainability practices within organizations.

7.2.5. Explainability System Constraints

Now that the *who*, the *why*, and the *what* are determined choices need to be made to develop the explainability system constraints. First, there will be certain trade-offs that must be determined and prioritized. After the trade-offs have been determined there will be certain system limitations, these are the elements that could not be incorporated within the constraints. Eventually, explainability system constraints can be finalized and determined that can guide the general system design and the design of explainability approaches.

	Stakeholder question	Explainability Design Constraint
1	the TM analyst asks why the TM model alerted an customer	- The feature importance method must provide selective features due to time constraints - The feature descriptions must be documented in the model documentation - The limitation of the explainability method must be well-known to avoid over-reliance - Communication between the developer and operator must be established for questions
2	The DNB wants to know how SIRA risks are reflected in the model	- The features must be linked to SIRA risks - The TM model output must be linked to SIRA risks to validate the reflection - The operator must be able to link the TM alert to SIRA risks - Model documentation must display the established reflection
3	The AP wants to know how the TM model handles certain types of data	- Model documentation must show the global behavior of data handling and processing - Model documentation must consist of transferable elements showing the data usage and processing - Model documentation must be updated periodically and checked for validity

Table 7.1.: Example stakeholder questions and explainability design constraints

7.3. Socio-technical Control and Validation

Once the explainability constraints are determined, these can be translated into system design features. The system design features will be developed and placed within the hierarchical control structure of the organization. The empirical study has shown that multiple risks arise when

explainability approaches are not revisited and controlled for validity. Results showed that for example, operational documentation can lead to an increased gap between normative procedures and established procedures. Or when model documentation is not revisited and controlled for over time this can lead to inexplainable design choices putting the useability of the model at risk.

To validate whether the design features in the system are still coherent and satisfying the explainability constraints hierarchical control structures must be instantiated. Hierarchical control structures are discussed previously in Subsection 5.2.3 and consist of control levels that have reference channels and measuring channels. An example of such communication channels is pictured in Figure 5.1. The downward reference channel enforces constraints on the behavior of lower level and the upward measuring channel receives operational feedback checking how effectively the constraints are being satisfied. Specifically, the measuring channels are often non-existent but crucial in dynamic complex systems. To validate the satisfaction of the implemented explainability constraints these institutional arrangements, such as feedback channels through measurement, must be incorporated and will make the system resilient to changes over time.

7.4. User-centered Operationalization of Explainability Taking on a Socio-technical Perspective

Now that the establishment of explainability design constraints and the approach to control and validate are discussed, these can be put together into an actionable method. The main part of the method, together with examples, is presented in Figure 7.1. The condensed method, together with the added element of socio-technical control can be seen in Figure 7.2. It is important to note that the method is an iterative process that provides system design constraints that must be controlled for. Each step within the method covers elements that are prone to changes and must be revisited periodically.

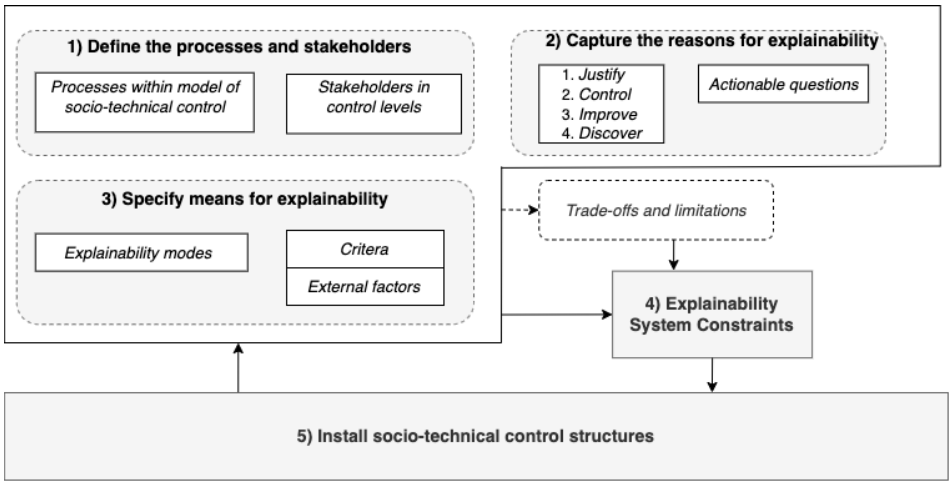


Figure 7.2.: Method for a user-centered operationalization of explainability

7.4.1. Demonstration and Evaluation of the Method

The method has been demonstrated and evaluated within the local practice at the bank. The focus group consisted of 11 employees which were provided a Toy Case together with the method presented in Figure 7.2. The Toy Case is presented in Appendix P and consists of a speculative scenario where a new CFT model must be developed. The focus group received an illustration of a ML system together with involved stakeholders and processes. Next to this, the focus group received an example of a control structure reflecting the processes active in the ML system. The focus group was assigned with the objective of applying the method to the developed Toy Case for a duration of 45 minutes where additional information and answers to questions are provided by the designer. The focus group consisted of practitioners that are well-known with processes in ML development and AML and CFT processes. The Toy Case has been introduced by providing background into explainability, systems theory, and system safety theory. The goal of the demonstration was to determine how the developed artifact can be used in one specific case. Next to this, the evaluation has been set out to determine how well the artifact can help stakeholders in operationalizing explainability from a socio-technical perspective. The extensive evaluation of the method together with possible improvements are presented in Appendix P.

The majority of the focus group experienced the method as an excellent invitation to introduce a new way of thinking about explainability. As one attendee stated that the method helped ”To think more thoroughly about explainability and not only as SHAP output”. Another attendee shared that the method is a ”Useful exposition of all the different components, invites a new way of thinking about explainability that enables actions. It would enable design processes both in general and per use case”. Another attendee shared that: ”As a process manager, it’s useful to have this method to refer to when engaging with stakeholders on explainability requirements. The constraints (downwards) and feedback (upwards) setup is easy to understand and helpful”. The method also provides discussion on dilemmas, as one attendee stated that ”Very concrete to create the method for explainability. Brings up a lot of questions and dilemmas”. The attendees described that the method has a clear structure due to the multiple isolated components and provided a clear structure for approaching explainability. Most attendees experienced working with the method as innovative resulting in new perspectives and questions that did not come to their attention before.

The attendees provided multiple limitations, especially on the theory behind the control structure. As stated by one attendee ”Perhaps additional guidance could be provided on how to instantiate the control structure”. Other attendees also suggested that additional guidance on how to establish the control structure would benefit the method. In addition to this, attendees suggested that elements could be made easier to understand or to provide additional examples of implementations. One attendee also noted that ”Who to ask the questions? As a data scientist I can only reason so much about some aspects of explainability”. This shows that the method should be used in a collaborative effort. These results also show that the method does not stand on its own and should be complemented with extensive guidance on concepts such as control structures, systems theory, and system safety theory. Next to this, the method could be simplified and additional examples must be provided to optimize the usability.

7.5. Main Findings Chapter 7

This Chapter started by describing that explainability is currently approached technocentric, ad hoc, and through loose efforts and is in need of a structured socio-technical approach that takes the audience into account. This approach is set up by learning from two main concepts existent in system safety theory i) developing constraints to control behavior and ii) implementing control structures to control and measure the satisfaction of the constraints. These concepts have been incorporated within the development of a method in order to answer research subquestion seven:

SQ7: *What method can guide the operationalization of explainability within ML based decision support systems, taking on a socio-technical view?*

The method is innovating the view on explainability by using decades of experience from safety engineering. This novel approach resulted in a method that operationalizes explainability by developing explainability system constraints and the instantiation of control structures that measure the satisfaction of these constraints over time. The method can be seen in Figure 7.2. The method has been demonstrated and evaluated within the local practice by developing a Toy Case that has been used within a focus group. The attendees, all employees at the bank, stated that the method provided a structured approach to explainability and assisted them in approaching explainability from a new perspective that give rise to dilemmas and questions providing them guidance for developing explainability requirements. The limitations of the method are that the method assumes knowledge of system safety theory and systems theory. This limitation can be solved by providing additional guidance on control structures or by giving additional examples.

8. Recommendation for Approaching Explainability

The following recommendations are a summary of the insights gained from the empirical study and will approach explainability from a socio-technical perspective. The only way that organizations using ML based decision support system can implement the proposed method is to prioritize explainability within their company policy and operating standards. Therefore, these are general recommendations provided for organizations on what is needed to actually implement an effective operationalization of explainability. The recommendations can guide practitioners and organizations or inspire regulators on how to approach explainability. The proposed recommendations are taking the strategies for AI system safety implications from Dobbe (2022), Leveson Lessons from (Leveson, 2011), and the proposed operationalization of explainability from Chapter 7. It is important to note that the recommendations stand on their own, but are all interrelated. The main recommendations are shown below and will be elaborated further within the Sections, for a full elaboration of the elements discussed within the recommendations Appendix N can be visited.

1. Embed explainability in the company culture
2. Create an explainability development plan
3. Operationalize explainability using a user-centered approach
4. Install structured communication channels
5. Avoid complexity and re-think the actual objective

8.1. Embed Explainability in the company culture

First and foremost, explainability should be approached as a system property that must be incorporated and controlled throughout all layers of the organization (i.e. system), starting with the culture and company priorities. Management must be aware that designing explainable systems will help the future organization and will pay itself off, especially as most systems are becoming increasingly complex over time. It is important to realize that there will be future explainability requirements, for example within the newly proposed EU AI act - the first law on AI by a regulator. Often management is expressing concerns for explainability, but to make a change the concerns should be translated into true priorities by allocating resources. Therefore, the following three practices should be established within organizations using ML systems: develop policies and procedures on how to approach explainability, detail the explainability goals and actions, allocate resources, and assign responsibility and authority.

8.2. Create an explainability development plan

Building upon the existing policies, procedures, goals, resources, and responsible actors for the explainability practices a structured approach should be developed to translate what is necessary to approaches that are fitting the purpose and determine how to implement this. This can be done by creating an explainability development plan, this will consist of three main elements: align mental models with the involved stakeholders, perform research on suitable methods for explainability, and research the technical limitations and hazards of the identified methods.

8.3. Operationalize explainability using a user-centered approach

Now that the policies and operating standards are established, there is a clear allocation of responsibility, the mental models are aligned, and the possible techniques and limitations are known the established method can be used. The method for a user-centered operationalization of explainability is pictured in Figure 7.2 and elaborately discussed in Chapter **??**.

8.4. Install structured communication channels

In order to measure the satisfaction of explainability constraints, align mental models, and provide feedback communication channels need to be established. Next to this, the empirical study has shown that providing uniform elements within documentation can help stakeholders understand concepts and minimize keyperson risk. Therefore, clear communication channels must be installed and uniform communication elements should be considered that can help company-wide communication of technical and explainability elements.

8.5. Avoid complexity and re-think the actual objective

Taking a step back and rethinking how explainability has even become an issue at all leads to the last recommendation. The unconstrained technical ability of ML to add features or develop combinations of models together with the desire to optimize performance within organizations can result in the design trap of creeping featurism. This is a concept describing the systematic tendency to add or expand a product with additional features making the system become more complex (Winograd and Woods, 1997). This is one of the main concerns raised by practitioners. By striving for performance optimization of the models, such as TM models, they become more complex and use more and more features resulting in greater complexity of both the model and the system as a whole. This results in an overall system where it is harder to test, provide explainability, audit, review, and maintain while costs are rising (Leveson and Weiss, 2009). Model developers need to refrain from complexity and must make hard decisions on the functionality of models while taking into account the effectiveness, explainability, and maintenance costs. Avoiding complexity results in designing for the exact model objectives, and stakeholder requirements, and keeping the explainability reasons of the audience in mind.

9. Conclusion and Discussion

The objective of this research was to address the existing gap in academic research on approaching explainability in AI systems that provide meaningful explanations for the intended audience taking on a socio-technical perspective. The research tries to address the research problem by situating explainability in a socio-technical context and by developing a method for a user-centered operationalization of explainability in ML decision support systems. The method has been established within the environment of Transaction Monitoring and has been demonstrated and evaluated by practitioners. The method can be used by practitioners to establish explainability approaches that are meaningful and can be controlled for over time. This Chapter will conclude the findings throughout the research and provide the main results, limitations, and contributions. The Chapter will close off with recommendations for future research possibilities.

9.1. Main findings

This research was motivated by the mass adaptation of AI systems in society applied to high-stake decision-making systems which have brought a remarkable set of challenges and concerns along. One of the main concerns is that AI systems are becoming increasingly complex and opaque to the extent that the reasoning behind the predictions is almost impossible to understand. When these systems are deployed they can have a tremendous impact on people's lives throughout all layers of society. There has been a societal urge for developing a greater understanding to avoid the adverse consequences of inexplainable AI systems. This is supported by the rising interest in research on Explainable AI and shown in the adaptation of explainability and transparency requirements within regulations. As described in Chapter 1, Explainable AI is still in its infancy and there are existing challenges and knowledge gaps that call for further research. However, as displayed in the research gap, these efforts are using a technocentric approach. Next to this, the current research on Explainable AI has not (yet) adopted the intended audience as the main driver for developing explanations. Explaining is a social process and can only be meaningful when considering the intended audience (Barredo Arrieta et al., 2020). Lastly, there has been little research on approaching explainability within local practices. To address this, an empirical study has been performed within the Transaction Monitoring department of a bank. Transaction monitoring systems use ML to support decision-making processes that try to detect money laundering and terrorism financing. The field of transaction monitoring is highly regulated and needs to adhere to a multitude of regulations on explainability, in addition to this they are obliged to detect certain risk patterns and do so by developing ML models. The empirical study showed insights into the difficulty of developing and maintaining explainability approaches in a complex dynamic system. The result of the empirical study has been summarized in five recommendations that can guide organizations in developing proper explainability practices.

The goal of this research is trying to get closer to defining explainability by first situating explainability in the socio-technical context and secondly by developing a user-centered method to operationalize explainability. These two main elements are captured within the main research question guiding this thesis:

What does explainability entail in the socio-technical context of Machine Learning based Transaction Monitoring systems?

9.1.1. Situating Explainable AI in the socio-technical context

Situating Explainable AI in the socio-technical context broadens the view on explainability as an algorithmic-centric problem to a socio-technical problem taking into account social, environmental, and organizational elements. The systematic literature review showed that there has been extensive research and development in explainability practices. These practices are mostly focused on the development of technical explainability methods that try to provide insights into global and local model behavior. The most well-known methods, however, show technical discrepancies and limitations. Explainability methods often do not accurately represent the internal model behavior and have difficulties handling complexity and feature interactions. Next to this, there is not yet a consensus on key concepts within literature such as explainability and interpretability. There has been research stating that current explainability practices do not take the audience for which explanations are intended into account. These limitations have been validated by the empirical study where model developers stated that technical explainability methods do not accurately represent complex model behavior. Also, the practitioners confirmed that often explainability methods are designed using the intuition of the developer and do not take the intended user into account. In addition to this, the literature review has shown that explainability is affected by latent dimensions such as the domain, user expertise, time available, and other environmental factors. This has been confirmed within the empirical study as practitioners provided insights into risks that occur because environmental factors such as time pressure and keyperson risk are not included in the design of explainability practices. This has resulted in flawed requirements, inadequate decision-making, over-reliance on technical methods, and the discarding of entire models. Although technical limitations of explainability methods are hard to resolve and can only be addressed by additional research, the empirical study has shown that taking into account the environmental, organizational, and technical elements can solve previously occurred risks. This shows that explainability within systems is depending on technical elements, social elements, and organizational elements. Therefore, situating explainability in the socio-technical context starting from the design phase is a necessity. When placing explainability in the socio-technical context these elements will be accounted for ensuring that external, social, and technical factors are taken into account.

By expanding this view on explainability approaches, that previously were not considered within explainability, are now included. These approaches vary from model documentation, operational documentation, global methods, local methods, and communication and feedback. Considering concepts from systems theory and system safety theory there are similarities observed between safeguarding safety and explainability within dynamic complex systems. The risks that have been observed within local practice resulting from explainability approaches are often due to misaligned mental models, missing validation of explainability with the end user, and unanticipated asynchronous evolution. Therefore, inspiration has been taken from system safety to establish design constraints that must be controlled for in order to establish explainability.

Explainability is hard to capture, however, when considering the elements that influence explanations covering the social, technical, and organizational factors while establishing control structures that measure the validity and understanding of explainability approaches provides a starting point for realizing explainable systems.

9.1.2. A method for a user-centered operationalization of explainability

Now that concept of explainability is expanded by considering the socio-technical context and viewing it as a control problem the main issue observed in practice can be addressed. Observed within literature and confirmed in local practice the intended audience is still missing within the design of explainability approaches. Next to this, the literature shows that there is a need for practical methods how to establish explainability. Explainability can be fulfilled when the requirements (i.e. reasons) of the intended audience are fulfilled. Because explainability is already considered an ambiguous concept and this research tries to operationalize explainability by developing a user-centered method. A more structured socio-technical operationalization can help practitioners approach explainability while being able to control the validity.

Learning from system safety engineering, explainability constraints must be established and enforced by using control structures to control the satisfaction of explainability over time. To achieve this, a method for operationalizing explainability has been developed taking on a user-centered perspective. This constitutes the eventual goal of the Design Science Research approach and will develop an artifact. The method is inspired by two main constructs of system safety theory which are (i) the development of constraints and (ii) the hierarchical control structure.

The user-centered operationalization of explainability, seen in Figure **??**, starts with defining the processes and stakeholders within the system. Next, the reasons for explainability are captured and turned into actionable questions that can be measured and realized. To address this, the explainability modes (i.e. approaches) that can aid in realizing the reasons must be listed together with the external factors that influence explainability. Eventually, this produces limitations and trade-offs which must be taken into account before defining explainability system constraints. As mentioned before, these constraints are imposed upon the system and must be satisfied by establishing control structures that ensure validity over time. Following the method is not entirely straightforward and should be revisited periodically, the method must therefore be seen as an iterative process. The method has been demonstrated and evaluated by using a Toy Case presented in Appendix P. Attendees from the focus group stated that the method helped them to establish a structured approach to explainability and that the method invites discussing questions and dilemmas between stakeholders. The attendees stated that the method should be extended by providing additional insights into the system safety engineering and particularly the hierarchical control structure.

9.1.3. Recommendations for approaching explainability

There is a need for broadening the view on explainability and practitioners must start to prioritize explainability from a system-wide perspective first in order to control lower-level behavior. There are recommendations provided in Chapter 8 that provide guidance for organizations that use ML based systems on how to establish a system-wide view on approaching explainability. The recommendations come forth out of the findings in the application environment and

specifically from the empirical study. Combining the empirical results with the theory of system safety has resulted in the following five recommendations.

1. Embed explainability in the company culture
2. Create an explainability development plan
3. Operationalize explainability using a user-centered approach
4. Install structured communication channels
5. Avoid complexity and re-think the actual objective

For an elaborate description of the recommendation Chapter 8 can be visited. The proposed recommendation takes a socio-technical view by addressing both technical recommendations, social, and organizational. The first recommendation is based on the main construct of system safety, which is that high levels in a hierarchical structure dictate lower-level behavior. So, first of all, policies and standards must be created in organizations to deal with explainability. And prioritization should be shown by allocating sufficient resources to be able to develop and execute explainability practices. The second recommendation shows the importance of aligning the mental models of stakeholders to gain a shared understanding of the capabilities of the ML system, only then effective requirements can be set up. Next to this, the state-of-the-art explainability approaches must be researched together with their technical limitations to discover the possibilities and potential challenges. The third recommendation is to structurally operationalize explainability from a user-centered perspective using the method designed in this research. The fourth recommendation states that structured communication channels must be installed to stimulate communication throughout the entire organization. This can help shape constraints and can result in improved policies and standards when lower-level operations can share their experiences. Also, uniform communication will aid in understanding documentation and will prevent the loss of knowledge with the departure of their creators. The fifth and last recommendation reflects on the usage of complex ML algorithms and models, this recommendation can be interpreted as a critical reflection to look at the system development and find out what the actual objective is. Often features or new technologies are adopted due to the promising result they might portray, but effectiveness should be the most important factor in considering technologies or algorithms.

9.2. Contributed Artifacts

The goal of a Design Science Research approach is to develop an artifact that contributes both science and alleviates a practical problem within the practical environment. During the process of developing the method for a user-centered operationalization of explainability, multiple artifacts have been created. Research from Offermann et al. (2010) categorizes artifact types that can be developed within Design Science, the artifacts created within this research are pictured in Figure 9.1.

Within the research two main artifacts are developed and examples are provided for a third. First, a pattern has been established by placing explainability in a socio-technical perspective together with applying concepts from system safety theory. A pattern provides generalized system design elements that can be used for many different kinds of designs (Offermann et al., 2010). This established pattern for explainability contributes mainly towards science and is

used as a language (i.e. building block) for the eventual artifact. The main artifact is the established method consisting of ordered activities that can be performed by people to support system development, in this research the method is focusing on operationalizing explainability within systems taking the audience into account. The method provides deliverables of activities for explainability and explainability system design constraints. The method is contributing towards science but is mainly contributing to the local practice, as has been stated by the attendees of the evaluation session. Using this method can provide a third artifact, which are requirements that make statements about the system. These requirements are reflected in the explainability system design constraints, examples for these constraints are provided in Table 7.1. The artifacts of requirements are not entirely produced within this research and do not hold as a contributed artifact.

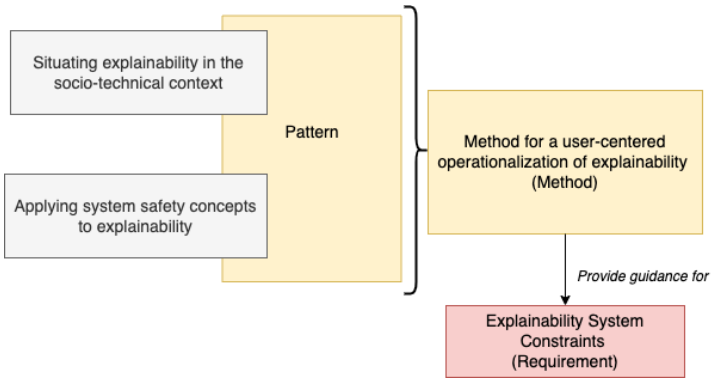


Figure 9.1.: Different types of contributed artifacts (Offermann et al., 2010)

9.3. Limitations

This section will discuss the limitations of the research to gain a better understanding of how to interpret and value the research results.

Research Methodology and Approach
Design Science Research has been performed which consists of three cycles: the relevance cycle, rigor cycle, and design cycle. The design cycle is an iterative process and should be tested and evaluated to use subsequent feedback to refine the design further. Although the method has been demonstrated and tested, there has not been an iteration in the design cycle to incorporate the retrieved feedback. Next to this, the demonstration and evaluation of the method provided limited results due to the amount of time and people provided for the focus group. The research approach consisted of two main elements, the systematic literature review, and semi-structured interviews. The systematic literature has been performed to gain insights into the body of literature on Explainable AI techniques and Machine Learning used in transaction monitoring. The literature review has been performed at the beginning of this research with the goal of defining an academic knowledge gap. In hindsight, an additional literature review specified on Machine Learning Systems and System Safety Theory could have provided additional value. This research has used a great body of literature on both, however, it could be beneficial to do so in a more systematic way. The semi-structured interviews sought to investigate the reasons for explainability, current explainability methods, and potential risks of explainability within the

local practice. One of the limitations is that 15 of the 16 interviews have been performed within the same department at the bank, this can limit the variety of perspectives on the addressed topics. Also, because this research has been performed at the bank the interview with the regulator provided limited information due to the danger of revealing information that can not be known at the bank.

Knowledge Base
The knowledge base consists mostly of research on explainability, systems theory, and system safety theory. Explainable AI has become a widely researched topic in the past few years. Most of the research, and papers used in the literature study, have been from the last five years. This brings the limitation of validation and consensus because explainability is such a new field there is still a lot to discover. Also, there is not yet a consensus on certain definitions (explainability vs. interpretability) or on certain views such as performance vs. complexity (myth or trade-off?). This can limit the current research in defining definitions or approaches that are wrong in hindsight or limited to the selection of literature viewing. System safety theory has already been applied to AI systems with the work from Dobbe (2022). However, approaching explainability as an emergent system property that can be viewed as a control problem has not been researched, until now. Applying system safety theory to explainability is a new concept that is not yet validated or peer-reviewed by other researchers. Therefore, it is strongly suggested in the recommendations for future research to further explore this novel approach.

Application Environment: Transaction Monitoring
The application environment was that of Transaction Monitoring at a bank which poses certain limitations. Transaction monitoring is highly regulated and has recently been under the great attention of both the regulator and the general public. This result in the fact that the actual risks, hazards, or applied techniques could have been communicated differently than reality. Also, money laundering is a highly secretive topic where criminals can not find out the exact approaches of transaction monitoring systems or typologies in order to prevent them from gaming the system by reading this research. Also, data could not be used and published for applying explainability methods on the data of the bank. All these factors posed limitations on the availability of data, access to information, and ability to present findings.

9.4. Recommendation for Future Research

This section will elaborate on several recommendations for future research, some coming from the limitations of this research and others from insights gained during the research which are in need of further research.

First, according to the design cycle from the Design Science Research approach the artifact created should be revisited to incorporate the feedback from the evaluation. Therefore, this research recommends an additional extensive evaluation of the method by using Toy Case or a case within a different domain, this will lead to insights on how to apply the method in practice and aid in optimizing the method. Also, it is recommended to further expand the scope to other banks to gain an additional variety of insights. Whenever researchers want to generalize the insights from this research even more it would be recommended to apply the method in other industries and within Machine Learning systems outside of supervised and unsupervised machine learning.

Another recommendation is to further research the view on explainability as an emergent system property. Additional research can validate the results in this thesis and potentially add valuable insights. Especially, developing control structures for explainability constraints is in need of additional research as mentioned by the attendees of the focus group. Another recommendation is to investigate the role of the human operator in the control structure and how the hands-on experience could improve the general Machine Learning system by researching the role of the human-in-the-loop.

Next to this, there is still a need for researching explainability from a socio-technical perspective and researching real-life use cases. Currently, there is a dominant technocentric approach that should be broadened. Insights from this research also showed that there is a need to further research evaluation metrics and methods for explainability techniques. Research to develop a comparison method on the properties of explanation methods and how to evaluate this would be valuable. In line with this research, it would be recommended to establish a general evaluation of the explainability of algorithms and how to determine how complexity is established. Insights from the empirical study showed that there is still a missing method that shows algorithms, their explainability, and how complexity is established. This can be applied in practice and especially regulators would be interested in the scoring of algorithms and models. Another interesting recommendation would be to make explainability more interesting for developers and to research how they can tailor explainability methods to built-in user requirements, such as more natural language explanations that can be provided to the intended end-user.

9.5. Personal Reflection on the Research Process

This research has taught me to develop a structured approach to analyzing scientific findings and translating these into workable concepts. Within the research a systematic literature review has been performed, however, throughout the process of the entire thesis, there has been a tremendous amount of literature reviewed. I have learned to analyze papers and select elements that can add value to my work. Next to this, I have learned to develop a structure for my own learning and work. In the beginning, this was hard to manage as there is no one else relying on my progress but myself. Also, I have learned to link elements from different research areas in moments of creative exposure. Lastly, I would like to mention that the process of working and presenting within a corporate environment as a researcher has taught me a lot. Within academics highly conceptual elements are often discussed, however, when presenting my findings at the bank I was forced to reduce these concepts to tangible components. This was hard but contributed to my analytical thinking. Structural analysis, work and time management, creative thinking, and communicating tangible concepts are the main learning's this research has taught me.

Although it is an unfair question, would I do it differently when starting over again? I can answer this with a definitive yes. I would try to begin with the end in mind. In my following research, I would try to picture and describe the concept I want to work on earlier and try to materialize this as soon as possible. I have learned that I can often be distracted by elements that are (really interesting) too specific and do not contribute towards my eventual goal. Next to this, I would increase communication with my supervisors and try to pose actionable questions. Especially in the beginning, there have been a lot of conversations about possibilities. However, next time I would try to start as small and concrete as possible and extend later on. These are the improvements I would make for my next research efforts.

9.6. Advice to the Dean of TPM

Now that I have completed my bachelor's and master's at the Technology, Policy, and Management (TPM) Faculty would there be elements that I like to change? Yes, but luckily most of these elements are being incorporated currently. During my bachelor's, I had a strong desire for additional technical courses which I eventually found in my minor and electives in my master's. I am excited to see the new courses on programming, ML, and AI being developed. However, there are two main elements that could still be incorporated. One of them is to incorporate the public and private domains into research assignments and projects. The interplay between the regulators and company management has interested me a lot, often during my time at TPM the main focus was on the institutional environment. However, broadening this scope to the institutional environment with companies could attract the interest of students. Next to this, it would be amazing to incorporate more use cases within practical assignments. This could be in collaboration with companies or not. During my master's course 'Fundamentals of Data Analytics', the assignments placed us in a fictional use case such as the fraud detection team of a bank that needs to find certain behavior. I have noticed that this fired up all the students. Another optional element could be, to prepare students for interaction with companies, is to incorporate optional elements of doing research at a company (such as MIP). I noticed that it helped my critical thinking a lot and even structured my academic reasoning. Lastly, I would like to say that incorporating new emerging technologies is making all the students extremely enthusiastic. This could be done within coursed or by inviting more speakers to the faculty. The introduction to Blockchain technology during my studies was amazing and I have learned a lot. Finally, I would like to add that when introducing such technologies many students are interested in learning (more) from the technical aspects, it could be amazing to provide (or reference) such information to scholars that want to gain additional knowledge.

A. Systematic literature review

The systematic literature review has adopted the search and selection strategy guidelines from Kitchenham and Charters (2007) and focuses on finding high-quality original research to identify, evaluate, and interpret the existing body of recorded documents on XAI and Detecting Financial Crime (DFC) with a focus on AML.

To find articles that are relevant to the research topic a search strategy has been designed, this is part of the literature scoping process that forms the foundation for the systematic literature review. The literature scoping process has been visualized in Figure A.1 and will now be elaborated upon. The search strategy starts with determining key terms that can be used in the subsequent search process. After this, the candidate articles will be filtered en selected using multiple selection criteria to identify the most relevant studies for the chosen domain. Lastly, the selected articles will undergo a final assessment to refine the existing body of articles, and this will result in a final selection of articles.

A.0.1. Search strategy

For the search strategy, the main search terms are derived from the major global components of the research topic, these can be seen in Table A.1 alongside the search combinations used. It is good to note that these are the search terms used in the eventual systematic literature review search, a preliminary search has been performed for explorational reasons. To find the relevant studies the two electronic library platforms Scopus and Google Scholar have been used.

For the retrieved articles, there were no geographical limitations or boundaries set for the year published. The only limitations set were that the writing must be in Dutch or English and that only the first ten pages of the retrieved outcome of the digital libraries are being taken into account. The search strategy was based upon finding the most relevant literature for XAI within decision-making processes using AI in the field DFC.

Keywords			
1)	Explainable AI	OR	XAI
2)	Fraud detection	OR	Financial crime
Search			
1)	(‘1’) AND (‘2’)		

Table A.1.: Main search terms and search combinations

The search phase resulted in 127 articles from Scopus and 100 articles from Google Scholar for search combination 1. Now that the search phase has ended the selection phase proceeds with 227 candidate articles.

A.0.2. Selection phase

The selection phase aims at applying more in-depth selection criteria on the candidate papers to find and select the body final of studies. The selection phase consisted of scanning the abstract of all articles to find out if the studies included one of the following criteria:

1. Touches upon definitions, potentials, and limitations within the field of explainable AI
2. Touches upon practical methods of explainable AI
3. Touches upon the use of AI systems for detecting financial crime
4. Touches upon the use of explainable AI for AI systems used for detecting financial crime

As can be seen in Figure A.1, reading the abstract of the candidate papers resulted in 28 papers from Scopus and 42 papers from Google Scholar that met the criteria. After eliminating duplicates within the selected papers there are 65 papers left that are going to be scanned entirely and assessed based on the same initial criteria.

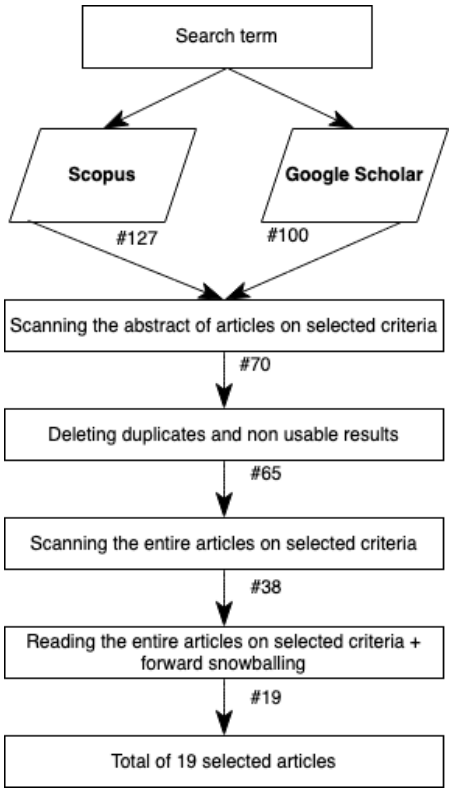


Figure A.1.: Literature review selection process

Eventually, a selection of 38 papers is read entirely and assessed, this resulted in a final selection of 16 papers that can be used for the literature review. An overview of the selected papers is given in Appendix B. The articles from the literature review have been thoroughly analyzed, for these articles forward snowballing is applied to potentially find additional valuable papers

Jalali and Wohlin (2012). Eventually, 19 papers are set out to be analyzed which will provide a solid foundation for the academic research gap.

B. Systematic Literature Review

Author(s) and year	Title	Criteria
Angelov et al., (2021)	*Explainable artificial intelligence: an analytical review*	1 and 2
Arrieta et al., (2020)	*Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI*	1
Bove et al., (2022)	*Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users*	1 and 2
Cirueira et al., (2021)	*Towards Design Principles for User-Centric Explainable AI in Fraud Detection*	2, 3 and 4
Chromik et al., (2021)	*I think i get your point, AI! the illusion of explanatory depth in explainable AI*	1 and 2
Coma-Puig & Carmona (2021)	*A Human-in-the-Loop Approach based on Explainability to Improve NTL Detection*	1 and 4
Das & Rad (2020)	Opportunities and Challenges in Explainable Artificial Intelligence (XAI)	1 and 2
De Bruijn et al., (2021)	The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making	1 and 2
Jesus et al., (2021)	*How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations*	1, 2, and 4
Kute et al., (2021)	*Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering – A Critical Review*	3 and 4
Meske et al., (2020)	*Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities*	1
Nesvijevskaia et al., (2021)	*The accuracy versus interpretability trade-off in fraud detection model*	3 and 4
Nicholls et al., (2021)	*Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape*	3
Psychoula et al., (2021)	*Explainable machine learning for fraud detection*	1, 2, and 4
Rudin (2019)	*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*	1 and 2
Sahakyan et al., (2021)	*Explainable artificial intelligence for tabular data: A survey*	2
Samek & Muller (2019)	*Towards Explainable Artificial Intelligence*	1
Sperrle et al., (2021)	*A Survey of Human-Centered Evaluations in Human-Centered Machine Learning*	1
Zhu et al., (2021)	*Intelligent financial fraud detection practices in post-pandemic era*	3 and 4

C. Research Flow Diagram

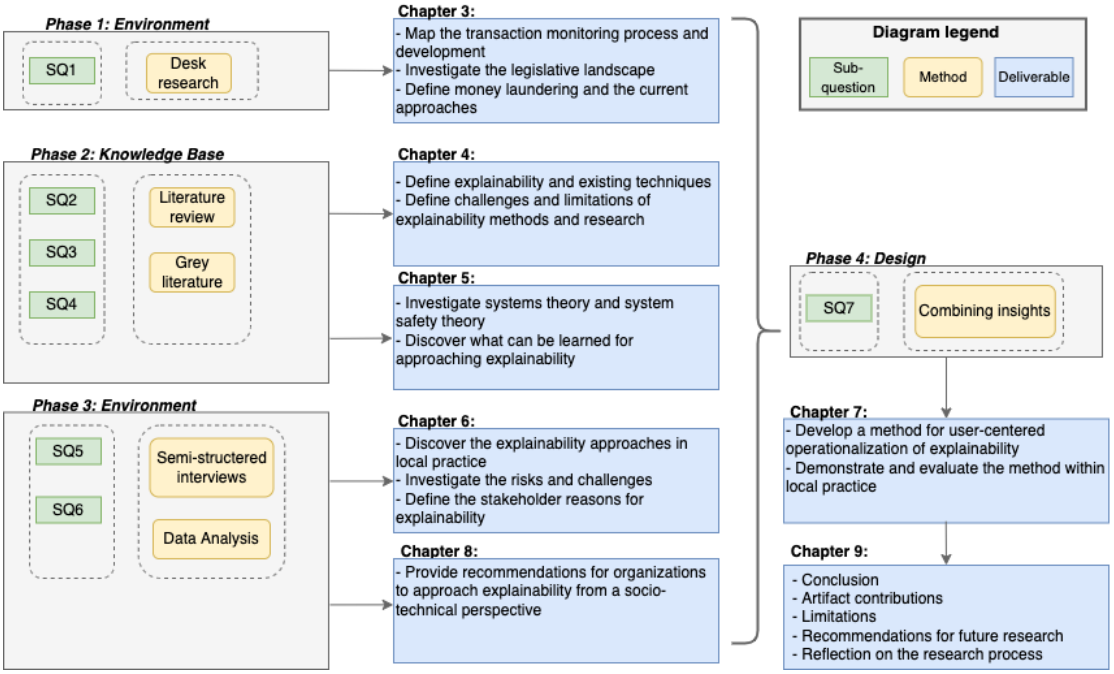


Figure C.1.: Research flow diagram

D. Artifact

Artifacts have four main components and are described by Johannesson and Perjons (2014) as the *function* of the artifact, the *structure* of the artifact, the *environment* of the artifact, and the *effect* of artifact on the environment.

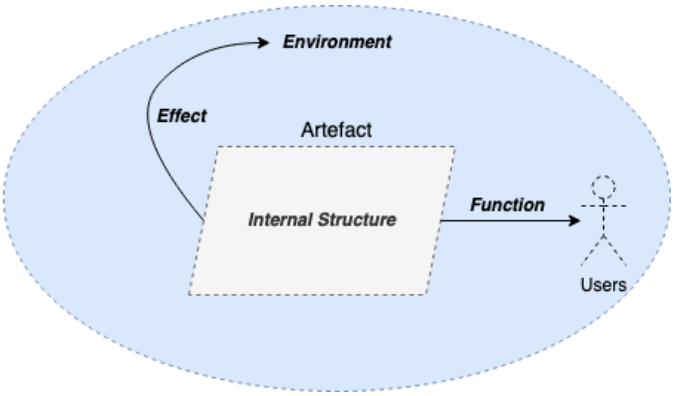


Figure D.1.: Artefact dimensions within design science research based on Johannesson and Perjons (2014)

The four dimensions of an artifact are depicted in Figure D.1. First, the artifact is defined by its structure representing the components, inner workings, and the relations between them. Next to this, the artifact has a function representing what it is designed to accomplish for users. Also, an artifact is always operating within an environment, this consists of all external aspects in which the artifact will function. Lastly, the artifact will influence and change its environment, this is the effect of the artifact. There are intended effects and side effects of the artifact.

Within DSR there are multiple ways of classifying artifacts, often this is done based on the type of knowledge they express or according to their function. As described by Johannesson and Perjons (2014) there are four main types of artifacts: constructs, models, methods, and instantiations. These types are shortly described below.

- Constructs: convey definitional knowledge and consist of terms, notations, definitions, and concepts needed for formulating a problem and the possible solution. Constructs enable structuring the understanding of phenomena.
- Models: express prescriptive knowledge and represent possible solutions to practical problems. Models prescribe the structure of other artifacts and consist of interrelated constructs.
- Methods: express prescriptive knowledge and define guidelines and processes on practical problems and reach certain goals. Methods prescribe how to create artifacts.

- Instantiations: exists of systems that are working and can be deployed and used in practice.

E. Processes of Money Laundering and Financing of Terrorism

Schott (2006) created an overview of these main three processes displayed in Figure E.1, these processes further are elaborated below.

- **Placement**: the initial stage is concerned with placing the resources in the financial system. The placement often happens through a payment service provider such as a financial institution. There are many different placement techniques, but, it may often involve a cash transfer of funds which can be divided into smaller less conspicuous amounts and placed during a certain time span over a multitude of financial institutions. Exchanging currencies or aggregating smaller notes into larger ones can also be used or even the use of converting or acquiring other financial instruments, such as securities, to divert suspicion (Schott, 2006).
- **Layering**: after the placement is successful the next stage is concerned with moving or converting the funds to other financial entities or instruments with the goal of concealing and separating the funds even further from the original criminal source.
- **Integration**: the final stage is concerned with getting the funds to the intended destination, for money laundering, this will be the legitimate economy and for terrorist financing, this is the designated terrorist payee or their supporting organizations.

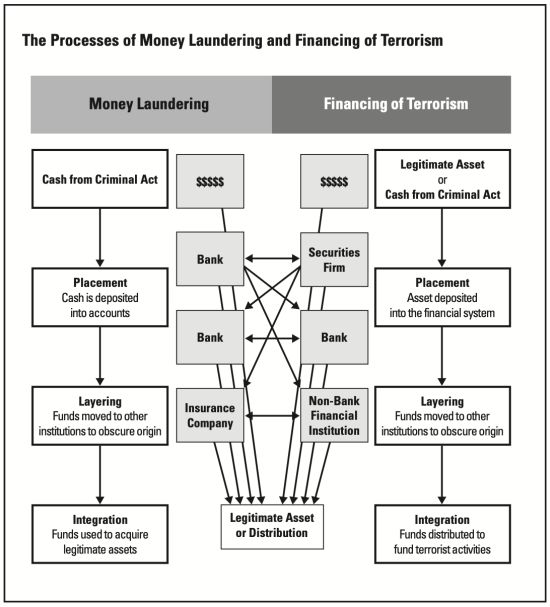


Figure E.1.: Processes involved in money laundering and financing of terrorism from Schott (2006).

F. Legislative Field for AML and CFT

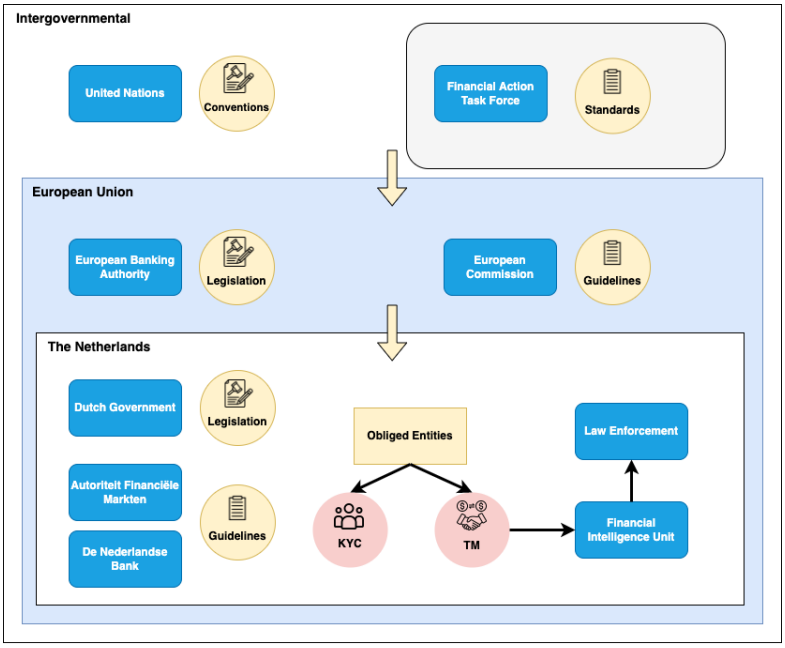


Figure F.1.: Overview of the key components of the legislative field for anti-money laundering and countering the financing of terrorism

G. Transaction Monitoring Process

The global process of TM consists of three steps that can be divided into monitoring (alerts), investigating, and reporting. These steps will be further elaborated down below.

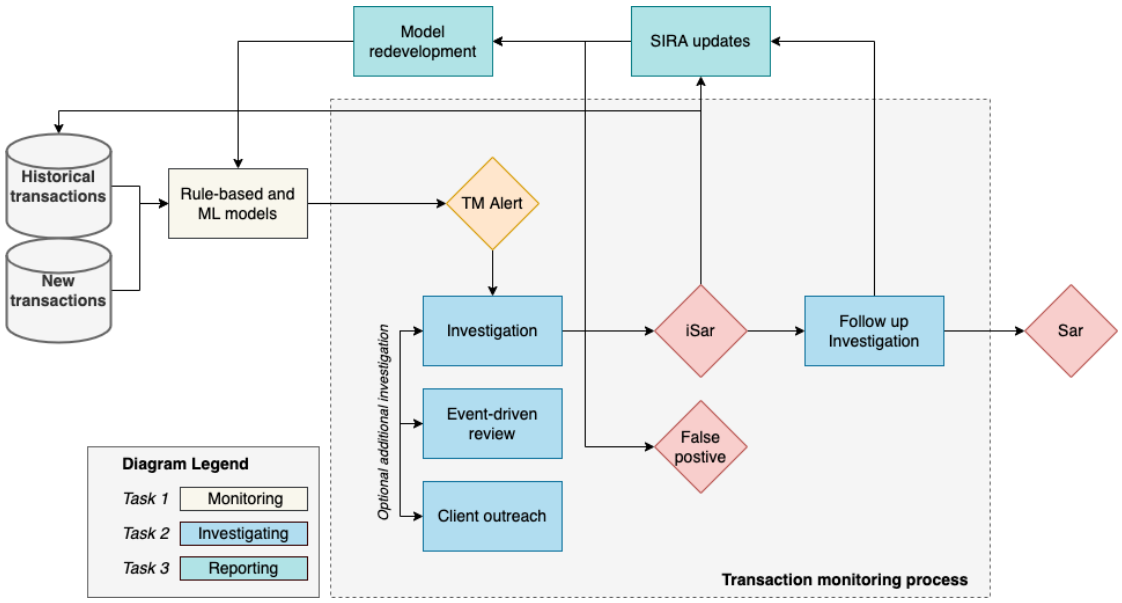


Figure G.1.: Process of transaction monitoring

Task 1 **Monitoring**: The first task is to monitor and detect fraudulent behavior as fast and precise as possible, therefore, this step is usually automated by a rule-based system that implements business rules to mitigate risk events from the SIRA that can lead to setting out an alert on a particular client or transaction. In addition to this, ML models monitor transactions and are trained on historical transaction data, the output of the model is a risk probability that is turned into an alert whenever a predefined threshold is violated. Eventually, both models determine unusual transfers of funds which leads to an alert set out on a specific customer.

Task 2 **Investigation**: The second task is to investigate the alert of the customer, which is performed by the TM analyst. The goal is to establish whether the transaction behavior can be explained and is low risk, or if no satisfying explanation can be provided and additional steps are needed. Within this step it is important to be able to justify the decision. However, it may be necessary, when the model output is not evidently pointing towards an obvious case of suspicion, to perform additional investigation measures. These additional investigation measures can be Client Outreach, where the client is approached for

additional information in order to gain a deeper understanding of their behavior. Another optional investigation method is the Event-driven Review, consisting of directly investigating other transactions of the client or involved parties to control and reduce risks. These investigations are initiated after an iSAR is filed and often take a long period to execute. The TM analyst determines whether these additional resources or actions are deemed necessary. The result of the initial investigation can be either concluding that there is no suspicion for money laundering or filing an iSAR in the case that there is a confirmed suspicion. After this, the iSAR is investigated more thoroughly by other second line parties who will determine whether a SAR will be filled and reported to the FIU.

Task 3 **Reporting**: The third task is to mitigate and report future risks by strengthening and further specifying the SIRA risk event library. These emerging risk patterns can be incorporated within the redevelopment of the model or a new model can be developed to mitigate these risks. A post-hoc analysis will be performed in order to improve the model and the detection process. Within this task, suspicious patterns or behavior will be evaluated and potential detection failures are optimized within the existing models. The investigation task can lead to new insights or confirm existing ones, either way, the knowledge gained should be used to increase the accuracy of the model. This can be done by simply adding the true suspicious alerts to the training data and by integrating new SIRA risks into corresponding features which can be included in the model redevelopment and/or new rule-base scenarios.

H. Transaction Monitoring System Development

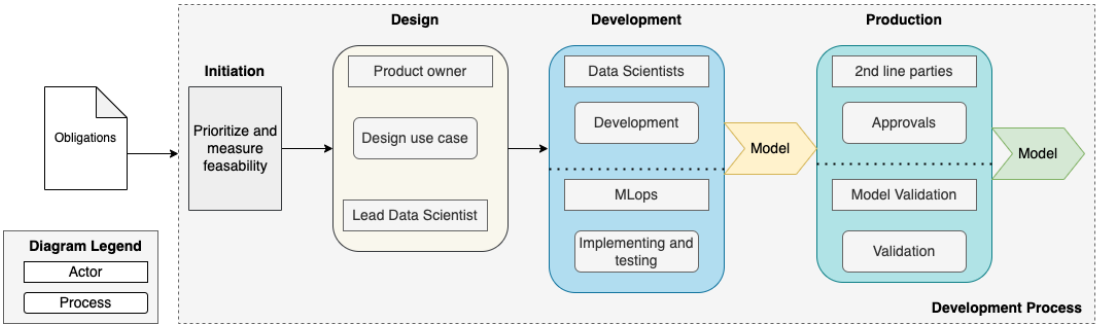


Figure H.1.: Process of developing transaction monitoring models

Step 1 **Initiation**: The first step of any project is often an idea, business incentive, or obligation that initiates the need for development. Within banks, this step is either started due to legislative obligations or obligations towards their customers for protecting them in more advanced ways or to different types of risks. The obligations and risks together with their impact and mitigating measures are prioritized using the RBA while taking into account the institution's risk appetite. Next to this, the feasibility should be determined taking into account the available resources, time, risk coverage, and legal aspects. After these steps have been performed the project ideas are turned over to the design phase according to their priority and feasibility.

Step 2 **Design**: Within the design phase obligations and ideas will be materialized defining concrete problem definitions which can turn into model requirements. The overall goal is to set the boundaries, constraints, and requirements for the use case. Within this phase often a Risk Owner takes ownership of the use case. Then, together with a Lead Data Scientist requirements are communicated and design concepts are established. Together they will dissect the problem into small manageable sections so that the development can be performed smoothly without miscommunication or misinterpretations.

Step 3 **Development**: When the use case is clearly defined the model can be built by Data Scientists, this process often consists of two phases where one is concerned with data handling and the other with the actual model building. Data handling is focused on data exploration, structuring the data, feature engineering, and all steps that need to be performed on the data so that it will suitable for the models (de Souza Nascimento et al., 2019). The other phase, in which the models are actually built, focuses on developing, training, testing, and evaluating the models. These steps often entail a feedback loop where stages can be revisited whenever goals are not yet met (de Souza Nascimento et al., 2019). After

the data and the models have been prepared and set up, they will be validated and, if they meet all requirements, accepted and brought into production. Within this phase, the focus is on operationalizing the model ensuring that the development, deployment, integration, testing, and releasing of the model all go well (Mäkinen et al., 2021). The model (re)development follows the MLops principles. After evaluation, feedback can be given to the Data Scientist on possible improvements or when performance is insufficient, this can focus on data handling or the model itself. When there is an agreement on all the specifics of the models and it is working within the environment the model can be brought into production. Finally, there is a testing phase which can be seen as a pilot of the model and processes. The model operates and sends out alerts to the TM analysts to see whether the desired results are generated, often there are multiple testing cycles in which each one is provided with feedback.

Step 4 **Production**: Now that the model is finished and validated the model can go into production, however, due to the fact that the model will be running in a corporate environment having an effect on real people impact assessments and approvals should be granted in order to go live. Privacy risks, and adherence to the GDPR, are assessed by Compliance. Also, Compliance investigates what can go wrong and what impact that would have, alongside this, they request mitigating actions for such risks from the development team. Model Validation assesses the technical workings of the model and checks whether there is a model risk. There are also checks for operational risks, residual risks, and general IT risks which are performed by a combination of 2nd line parties. Whenever appropriate adjustments have been made and the approvals have been granted the model can go live into production.

I. Interpretable Models

Linear Regression Models
Linear regression models are concerned with the task of regression modeling the dependence of a regression target (Y) on the set of features ($x_1 ... x_k$) which is often represented by a linear relationship:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

The obvious advantage of linear regression models is that the produced weights ($\beta_1 ... \beta_k$) of the features can directly quantify the influence of a feature in predicting the target outcome while taking into consideration the entire set of features. This can help in understanding the relative importance of different features. The first weight β_0 is the intercept and is not multiplied with any feature. Interpreting the weights of the features depends on the type of corresponding features, for the following example numerical features are assumed. The intercept shows the prediction value when all other features are at their reference value, for numerical features this is zero. However, often this is not relevant as instances with all features at zero do not make sense in practice. But when all features are standardized the intercept will reflect the predicted outcome of an instance with all features at the mean value. The weights (β_k) of the features can be interpreted as the change in predicted outcome when the features are increased by one unit.

The linearity assumption, assuming linear dependence of the predictors and predicted variables, ensures that linear regression models meet all three characteristics of transparent models. Linear regression models will always fall within the algorithmic transparency but when the model significantly increases in size the simulatability can be violated when humans can not think of the model as a whole. Also, when highly engineered complex features are used the decomposability can be violated. Linear models can also be interpreted by using techniques such as visualization in order to gain a thorough understanding of the feature importance or to some degree the interactions among them, this can be helpful for non-expert audiences. Post-hoc visualization techniques will be further discussed in Section 4.3.4.

It can be analyzed whether linear models can produce explanations that cover the description of a good explanation from Section 4.2.4. Linear models are not selected by default but this can be manually achieved by using fewer features or by training sparse models. Linear models are contrastive, although the reference instance is a meaningless instance that is unlikely to occur in reality or in the data (Molnar, 2020). Linear models do create truthful explanations whenever the linear equation is also accurately representing the relationship between features and the outcome. Whenever there are many non-linearities or interactions present the linear model will not be able to capture the actual relationships resulting in a less accurate and truthful explanation.

Logistic Regression Models
Logistic regression models are concerned with the task of classification modeling the dependence of a target (Y) through a linear combination of the given feature set ($x_1 ... x_k$). Logistic

Property	Assessment
Expressive Power	The linear regression weights can be directly interpreted by the correlation coefficients.
Translucency	High, all internal parameters can be accessed.
Portability	Low, explanations entirely rely on the inner workings of the linear regression model.
Algorithmic Complexity	Low, there are no complex methods used for generating explanations.

Table I.1.: Properties of linear regression explanation method

regression uses a logistic function that maps the linear combination on a probability interval [0,1], this mapping is represented by:

$$P(Y = 1) = 1/(1 + (exp - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)))$$

However, unlike linear regression, there is no direct mapping between the weights and the effect on the outcome as there is no linear relation. This is because the weights appear inside the exponential function so the importance of the features can not be directly extracted from the weights resulting in a loss of straightforward interpretation. With linear regression β_k is the relative importance of feature x_k and when interpreting the change of one unit in the feature causes the target to change by the weight (Gianfagna and Di Cecco, 2021). With logistic regression one unit change in the feature will change the odds by a multiplicative factor $exp(x_k)$, this implies that a probability for a certain class will be enhanced by the factor $exp(x_k)$. Such interpretation feels less intuitive and requires some sort of statistical knowledge.

Logistic regression models also fall within the three characteristics of transparent models. But similarly, the linear regression models must be controlled for size in order to meet simulatability and for complex features and interactions to adhere to decomposability. The logistic regression model can be understood by using visualization techniques which will be further elaborated in Section 4.3.4.

Logistic regression shares most of the advantages and shortcomings of linear models as well as similar characteristics of good explanations. Logistic regression has a great advantage over other classification algorithms that only provide a final classification as it outputs a probability for a certain class, this can be interpreted and can be useful in academic fields or high-stake decision-making areas.

Decision Trees
Models can exploit a graph structure taking the form of a tree. The tree structure starts with a root note containing the entire dataset. The dataset is split according to certain cutoff values in the nodes, by splitting the data multiple subsets of the dataset are created while adding data instances to these subsets. Eventually, when all splits are performed final subsets are formed at the end of the tree which are called leaf nodes. The intermediary splits, or subsets, are called internal or split nodes. Subsets are determined by a certain cut-off point that tries to make the resulting subsets as different as possible with respect to the target outcome. The tree algorithm continues to split until a certain stop criterion is reached. Edges between nodes can be viewed as 'AND' structures and the subsets can be viewed by inspecting them. Eventually, each data point is assigned to a leaf node which represents a predicted outcome for the subset of instances.

Property	Assessment
Expressive Power	The coefficients are not as easily interpretable as linear regression requires some specific knowledge.
Translucency	High, all internal parameters can be accessed and weights are used for explanations.
Portability	Low, explanations entirely rely on the inner workings of the logistic regression model.
Algorithmic Complexity	Low, there are no complex methods used for generating explanations.

Table I.2.: Properties of logistic regression explanation method

Trees can be used for both classification and regression. In order to predict the outcome in the leaf nodes the average outcome of the training data within the node is used or in the case of classification the majority vote. Various algorithms can be applied for tree structures and can differ based on the number of splits used, the criteria for determining the splits, and stopping rules for splitting (Molnar, 2020). There are also algorithms that develop multiple trees and learn them on different subsets of the training data after which the predictions are aggregated, these are called tree ensembles.

Tree algorithms follow an intuitive structure and are naturally good for visual interpretation. The adoption of decision trees in supporting decision-making processes comes forth out of their off-the-shelf transparency (Barredo Arrieta et al., 2020). Unlike linear model families, tree algorithms can capture interactions between features and data well. Also, the data points are easier to understand because they are all allocated in distinct groups, the leaf nodes, instead of interpreting the data points on a multi-dimensional hyperplane for the linear model families. Simple decision trees are simulatable but can become decomposable when the size of the tree and the number of features increases. The previous characteristics are lost when the size is further increased and complex feature relations are introduced making the decision tree only algorithmically transparent.

Decision trees generally provide good explanations following most criteria described in Section 4.2.4. Counterfactual explanations can be naturally extracted and determine in what scenario an instance could belong to other subsets. These what-if scenarios help in comparing predictions of instances by comparing the split points up until the leaf nodes. The selectiveness of the tree depends on the depth, i.e. the longest path between the root node and the leaves. When trees are structured with a small depth it becomes easy to understand and explain because there are not that many splits which divide the instances into the distinct groups, each of the splits are often binary decisions making them very well interpretable. Just like the linear model families, the truthfulness depends on the predictive performance of the tree algorithm.

Property	Assessment
Expressive Power	High, reasoning follows natural human-friendly way of explaining.
Translucency	High, all internal parameters can be accessed.
Portability	High, a lot of algorithms built upon the decision tree structure such as Random Forest or boosted trees. So, the results can be incorporated into these models.
Algorithmic Complexity	Decision trees are NP-complete but heuristics are used for optimization.

Table I.3.: Properties of decision tree explanation method

J. Additional Model-agnostic Methods

Visualization

Visualization techniques are most often applied to supervised learning models and are able to visualize representations of the model to explore patterns and detect relations. Visualization is the most human-centered interpretability technique and can produce good explanations for opaque model behavior, however, visualization techniques can also produce visually interesting visualizations that are not entirely understandable by the human (Adadi and Berrada, 2018). One of the methods used to realize visualization is to build surrogate models which will be described below.

Surrogate Models - LIME

A surrogate model is a simple interpretable model which is trained on the predictions of the original model in order to explain the latter (Adadi and Berrada, 2018). Global surrogate models are interpretable models that are trained to approximate the predictions of the black box model. The global surrogate model method is flexible and intuitive, and it can be easily measured how good the surrogate model is in approximating the predictions (Molnar, 2020). However, global surrogate models can fall short as they may not model global complexity, and feature interaction, and can reflect their own structural biases (Kaya, 2022). Also, it is important to be aware that conclusions are made about the model and not the data.

It often occurs that data lies on globally non-linear but locally linear manifolds, in order to leverage this there are also local surrogate models. Local surrogate models are used to explain either group instances or individual predictions of the black box model. The idea of local surrogate models is to understand why the underlying model makes a certain prediction by testing what happens to the prediction when a variation of the original data is given to the model. The most well know the local surrogate method is LIME and does so by generating a new dataset consisting of permuted samples and the corresponding predictions of the black box model. The new samples are generated by perturbing each feature individually, drawing from a normal distribution with mean and standard deviation taken from the corresponding feature (Molnar, 2020). LIME then trains an interpretable model, such as a linear model or decision tree, on the new dataset. LIME produces an interpretable model that should be a good approximation of the original ML model, which is not necessarily a good global approximation but focused on local fidelity (Molnar, 2020). LIME focuses on instance-wise predictions which are shown to support both domain experts and non-experts on model selection, assessing trust, improving untrustworthy models, and getting insights into predictions (Kaya, 2022). LIME is portable and when using specific interpretable models such as a decision tree the explanations are selective and possibly contrastive, resulting in human-friendly explanations. This is one of the reasons why LIME is a good method to inform laypersons or users with time constraints. However, LIME can not produce complete attributions which might be necessary in compliance scenarios or for the purpose of debugging a model. But concerns raised from Molnar (2020) state that there can be instability of the explanations when sampling the data multiple times which should be investigated before applying LIME. Also, LIME is based upon the assumption that ML models exhibit linear behavior locally which is not built on proven theory (Molnar, 2020).

Property	Assessment
Expressive Power	High, for single predictions and human-friendly.
Translucency	Medium, do not have extensive insights into the model internals.
Portability	High, the method does not rely on the inner workings of the ML model.
Algorithmic Complexity	High, computation time increase significantly and must be defined beforehand. A compromise must be made between fidelity and sparsity.

Table J.1.: Properties of Lime explanation method

Example-based Explanations

Example-based explanation techniques use particular instances of the dataset to explain the model's behavior or to explain the underlying data distribution. Example-based explanations are model-agnostic because they can make any ML model more interpretable but do not perform any kind of transformations or access the features such as other model-agnostic techniques. Because example-based explanations rely on the data they can only be leveraged when the data itself can be represented in a human-understandable way, this is for data where features carry contexts such as for images or text. Depending on the structure of the data and the number of features example-based explanations can also make sense for tabular data. Example-based explanations follow the natural human way of explaining decisions and are inspired by the cognitive science of human reasoning as this is often prototype-based using representative examples as a basis for categorization and decision-making (Kim et al., 2016; Adadi and Berrada, 2018; Lipton, 1990). This can help humans construct mental models of the model and the data or can improve understanding complex data distributions (Molnar, 2020). The idea behind example-based explanations is to learn from previous events and follows the general reasoning of Event B shares similarities with event A, A resulted in Y, so we could derive that B will imply Y. Next, the counterfactual explanation method will be elaborated.

Counterfactual Explanations

Counterfactual explanations can show how the model makes its predictions and can explain individual predictions by telling how an instance has to change to significantly change its prediction (Molnar, 2020). Counterfactual explanations use hypothetical reality to contradict an observed fact, they describe a causal situation in the form: "If event X had not happened, Y would not have happened". This hypothetical reality is a predefined output different from the original prediction. Knowing when and how a prediction changes in a relevant way can give insights into specific predictions and global model behavior. Taking an example for a loan applicant, counterfactual explanations can show what the smallest change to the features (age, income, debt..) is that would change the eventual prediction from the initial predicted rejection to an approval prediction. An important criterion that should be matched is that the counterfactual should be as similar as possible to the instance regarding the feature values and change as few features as possible (Molnar, 2020). Also, the new feature values should represent likely or realistic scenarios.

The interpretation of counterfactual explanations is very clear as they exhibit properties of human-friendly explanations because they are contrastive and selective. Also, counterfactual explanations do not rely on additional assumptions such as LIME and are relatively easy to implement. The counterfactual explanation method is extremely portable and can work with

rule-based models or other logic systems. As counterfactual explanations do not need to access the model or data and only the prediction function they can work with proprietary models. This will protect the interest of the model owner by offering explanations without disclosing the model or data. A disadvantage of counterfactual explanations is that there is not always a counterfactual instance found, this depends on the data. Also, when features are categorical each combination of feature values should be explored which leads to a computational explosion, however hopefully this may be improved in the near future by using an optimizer combining continuous and discrete inputs (Gianfagna and Di Cecco, 2021; Molnar, 2020). A final inconvenience is that counterfactual explanations often produce multiple counterfactual explanations for each instance, this can be advantageous as human can select the ones according to their domain knowledge but can also lead to practical challenges such as time constraints.

K. Traditional Safety Engineering Efforts

K.1. Traditional Safety Engineering

In general, systems are designed to fulfill certain goals while satisfying specific requirements and constraints. Engineering is a way of organizing the design process, which can be applied to systems, and tries to do so with the most cost-effective result. Safety engineering is concerned with designing systems that fulfill the requirements and constraints so that the system fulfills an acceptable level of safety. Verma et al. (2010) describes safety as a combination of reliability and consequences, whereas a system and its components should be reliable but also consequences should be reduced by providing safety control systems which anticipate failures minimizing their consequences. The origin of safety engineering goes back centuries, however, defined structured approaches for designing safe products and systems arose in the postwar era. During this time period, societal concerns and public debate opened up conversations on the safety of nuclear power, civil aviation, the development of lethal chemicals and weapons, and increased environmental pollution (Leveson, 2003). There was a growing need for ensuring safety for systems that can cause hazards such as human loss or injury but also the destruction of property and environmental harm (Leveson, 2003). The engineering approaches developed at the time were totally different and shaped for particular industries. The classical approach for safety engineering was based on analytical reduction, isolating system components, and applied to event-based accident models.

To illustrate the aspects involved in safety engineering an example within the civil aviation industry will be elaborated upon. Within the aviation industry, the approach to safety was built upon hazard identification linking accidents with specific aircraft components. Components are designed and manufactured using a fail-safe design with a high degree of integrity based on the reliability rate of the fault hazard analysis (Leveson, 2003). Aviation had a clear fly-fix-fly approach learning from the past and reiterating the process of the fault hazard analysis with the redesign and modification of components. For aviation, this approach was succeeding partly due to the fact that the commercial aircraft industry is conservative in its design approaches and with the introduction of new technologies. Institutions aided in the design of safety by setting up tight regulations for the commercial aircraft industry. However, the approach was not working perfectly when more drastic technological advancements were implemented resulting in increased accident rates (Leveson, 2003). An example of this was when glass cockpits were introduced in light aircraft aviation, the glass cockpits refer to the use of computer screens for pilots rather than the analog system and changed the way pilots monitor information in the cockpit. The glass cockpit placed greater demands on pilot attention and risked overloading the pilot with more information than they could effectively monitor and process (National Transportation Safety Board, 2010). The new cockpits relied on computerized systems integrating multiple data inputs increasing the complexity and potentially limiting the pilot's ability to understand the functionality of the underlying system (National Transportation Safety Board, 2010). The introduction of this new technology brought a new set of potential safety concerns

such as pilot performance, training, and new accident investigation techniques (National Transportation Safety Board, 2010). Such innovation causes pilots to make different types of errors and change the accident mechanisms of the earlier defined safety approach (National Transportation Safety Board, 2010; Leveson, 2003). The result from the research of National Transportation Safety Board (2010) on reviewing accidents involving light aircraft equipped with these glass cockpits found that a pilot's experience and training in conventional cockpits do not prepare them to safely operate the more complex glass cockpit. The research showed that there was a lack of information provided to pilots about the new cockpit itself, resulting in misunderstandings or misinterpretations of system failures. Concluding remarks from the research stated that the newly introduced glass cockpits did not improve safety when compared to conventional cockpits. The analysis identified two main safety issues and addresses the need for pilots to have equipment-specific knowledge and the need for capturing maintenance and operational information (National Transportation Safety Board, 2010).

The main takeaway from this example is that technical improvement, often data-driven and computerized may intuitively seem to improve safety. Achieving the potential safety benefits, and eventually, efficiency, will not be reached without proper procedures, guidance, and training.

K.1.1. Traditional Accident Models

Approaching safety engineering is often done by an accident model or hazard analysis used to determine and improve the reliability of components, processes, or fail-safe mechanisms. The accident models provide a means for understanding phenomena and are used to explain how accidents occur which can lead to the (re)design of safety control structures. Also, accident models are often used to find the root cause of the event to assign blame for the accident. Although it may seem that the main focus is on accidents that have occurred in the past, a good accident model can often be used preventively to improve the design of safe systems. Leveson (2002) describes that accident models '*form the basis of (1) investigating and analyzing accidents; (2) designing to prevent future losses; and (3) determining whether systems are suitable for use by assessing the risks associated with an activity, the use of a product, or the operation of a system*.

Traditional accident models explain accidents with a root cause that is followed by subsequent events all leading towards the accidents, the events taken into account almost always involve some type of component failure or human error (Leveson, 2002). When taking into account multiple events leading toward accidents, this is often captured in an event-based model or so-called event chain. The causal relationship between events, within an event chain, is direct and linear meaning that one event triggers subsequent events to happen and can be traced back to the root cause. The root cause often represents an explanation of the accident. This shows the nature of event-based modeling being focused on root causes that are nearby the accident and often does not take into account that the foundation of accidents can be laid down years before (Leveson, 2002). History has shown that whenever human operators are in the system they often become the root cause of an accident in event-based modeling.

Until the 1950s the focus of accident models was on the human operator, or in the example of aviation on the pilot, and was most likely to blame for occurring accidents (Leveson, 2002). During the specified postwar era aircraft accidents sky-rocketed whereas 8547 people died in the United States from 1952 to 1966 (Leveson, 2002). During this period a new approach to safeguarding safety shifted the focus from operator error as the cause of accidents to viewing safety

as a design characteristic. This new approach, introduced by H.A. Watson in 1961, argued that safety must be integrated into the design just as performance, stability, and structural integrity (Leveson, 2002; Verma et al., 2010). The Air Force only started considering this approach when autonomous intercontinental ballistic missiles were developed and no pilots were to blame for accidents (Leveson, 2002). Since then, the Air Force began to treat safety as a systems problem, introducing the evaluation of the system as a whole in safety engineering within the aerospace industry.

L. Leveson Lessons

The models used in traditional safety engineering do not include subtle factors, such as social and organizational interactions, that do play a significant role in system failures. Hence, research from Leveson (2002) suggests that the traditional accident models should be extended based on five dimensions, making them more effective for the emerging hazards in the complex systems of today. The dimensions in which the event-based models need extension are: the social and organizational factors, system accidents and dysfunctional interactions, human error and decision making, software error, and adaptation. Working along these dimensions, which stretch the limits of traditional safety engineering, seven new assumptions are developed as the basis for a new foundation for safety engineering in complex socio-technical system. The assumptions from Leveson (2011) are coined by Dobbe (2022) as *Leveson Lessons* and can be grouped into five dimensions: social and organizational factors, system accidents and dysfunctional interactions, human error and flawed decision making, software errors, and adaptation.

Social and Organizational Factors

In order to prevent accidents in complex systems the accident model must include both the social, organizational, and technical factors, only then the system can be completely understood and accidents can be effectively managed. The social system includes the purpose, goals, and decision criteria used to construct and operate the systems (Leveson, 2002). Next to this, the organizational factors play a major role in shaping interactions and the organizational structure, management, procedures, and culture of the engineering organizations that created the system should be incorporated in the accident models (Leveson, 2002). Traditional accident models do not represent systemic accident factors on a organizational level, taking into account potential deficiencies in the company culture, management practices, and safety culture of the company or industry (Leveson, 2002). Accident models should look beyond the proximate events of an accident and broaden the focus from merely technical components and pure engineering activities to the social system and organizational factors overlying the complex system to attain an acceptable level of risk control.

Multiple researcher proposed models to incorporate the causal factors of accidents on different levels of abstraction, whereas the first level describes the accident mechanism using an event chain, the second level shows the conditions that led to the events in the first level, and the third level is made up of the systemic factors that contributed to the accident incorporating technical, human, managerial, organizational, and societal facets (Johnson, 1980; Leveson, 1995, 2002). The model of socio-technical system involved in risk management from Rasmussen (1997) is the most inclusive, applying these hierarchical add-ons to event chains (Leveson, 2002). Within this model the social and organizational aspects are included using a hierarchical control structure, having levels for government, regulators, the company, management, and employees while defining the information flow between each entities.

Lesson 1: ***High reliability is neither necessary nor sufficient for safety.***
The first lessons adopts the notion of taking into account the social and organizational factors,

this extend the traditional safety engineering approach that limits itself focusing on component reliability. Previously, the idea was that when technical components have high reliability they assure safety. However, in the more complex systems of today this can not be supported anymore. The view should be broadened to factors that shape human behavior and take into account the social context.

Lesson 7: ***Blame is the enemy of safety. Focus should be on understanding how the system behavior as a whole contributed to the loss and not on who or what to blame for it.***
The final lesson is on the safety culture within organizations, the only way to apply adoptation and built resilient safety defence systems is to be able to learn and progress throughout time. Instead of focusing what an operator did wrong it is important, in order to prevent future accidents, to investigate why the operator made certain decisions under the conditions he was in. Only with a just safety culture this information can be gathered and people will feel comfortable to share their intentions

System Accidents and Dysfunctional Interactions

Traditional system accidents often occurred due to the failure of individual components which could not satisfy its specified requirements. However, when dealing with complex systems new system accidents arise concerned with the interaction among components. In such cases individual components might satisfy their specified requirements but the effects of interacting components might cause hazards for the system as a whole. Nowadays systems are interacting with other physical systems, humans, or software and dysfunctional interaction among system components can occur resulting in accidents. Such system accidents can occur due to the inability to thoroughly analyse and test all interactions in complex systems, resulting in inadequate control over the interacting components. Dysfunctional interactions should be identified, reduced, or eliminated to prevent interactions that can lead to a hazardous state in the controlled process (Leveson, 2002).

Digital technology and software driven systems increase the interactive complexity as well as increasing the coupling between components, resulting in more system accidents. Next to this, there is a growing need for high efficiency and functionality within these systems resulting in tightly coupled systems that do not allow for intervention when problems arise and can lead to cascading subsystem failures. Leveson (2002) advocates for applying systems engineering to deal with dysfunctional component interaction and emerging hazards by analyzing errors in the system design rather than merely focusing on the component design. This approach focuses on tracing the system functions to the individual components and classifying the types of dysfunctional interactions leading to accidents. Also, decoupling or loosely coupled components and subsystems can increase the ability to intervene but this is hard to achieve as society is trying to keep up with managing the fast increased complexity of systems today.

Lesson 2: ***Accidents are complex processes involving the entire socio-technical system. Traditional event-chain models cannot describe this process adequately.***
The second lesson shares similarity with the first lesson but is more focused on accidents. As systems have become more complex, the interaction among components have increased resulting in more complex system accidents and dysfunctional interactions. Whereas traditional safety engineering traces accidents down to individual component failure while defining a root cause in a linear event-chain, this is not possible anymore in the systems of today. The component interaction in systems nowadays are between physical, human, and software based components each having their own fallacies. This can result in system accidents and dysfunctional interactions, hence, the system design together with the interactions should be analyzed rather than individual components.

Lesson 3: ***Risk and safety must be best understood and communicated in ways other than probabilistic risk analysis.***
The third lesson focuses on shifting the view of accident and safety analysis from a mathematical viewpoint to a system design perspective. Traditionally, risk information is communicated in the form of probabilities and is most often only considering physical failures. However, such computational tools have serious limitations and accidents models should not be based on failure events. An example of a proposed accident model looking beyond probabilistic risk analysis is the System-Theoretic Accident Model, this model will be discussed in later sections.

Human Error and Flawed Decision Making

Human error, or operator error, is found to be the cause of 70-80% of the accidents and is often selected as the root cause in an event-based accident chain because there exists a deviation from the performance of a specified sequence of actions (Leveson, 2002). However, this can be misleading as deviation from a standard almost always tends to be true in practice when operators strive to increase efficiency and productivity. This has become so ingrained in industries, and even expected from organizations, that a common way for operators to set up a strike is to work to rule. This implies that the operators follow all prescribed sequences of actions and are using this as a threat to apply pressure on management. Operators, working under time constraints and pressure, may eventually define their own set of rational behavior deviating from the formal rules which can be labeled as established practice, or the activities that have been established over the years of working. The established practice follows the most effective procedures and can deviate from the normative work instructions and rules. This can result in a conflict when there is a need to determine whether an error or wrongful action has occurred, is this by deviating from the rational and normalized effective procedures or from the normative work instructions and rules? Often the latter is the case as established practice is in principle violating formal rules, defined earlier as human error. This shows the ambiguous nature of human operations.

Leveson (2002) argues that an important tool to align the established practice and specified practices, following the normative procedures and rules, is to align the mental models of designers and operators. A mental model is an overarching term for any framework, worldview, or concept a person carries in their mind to be able to understand phenomena. Mental models are naturally evolving through interaction with the specified system and are constrained by a person's technical background, experience with similar systems, and ability to process information (Gentner and Stevens, 2014a). Mental models differ per individual and are based on generalization and analogies from experiences, but mental models can also be defined for certain roles when setting up certain general assumptions (Gentner and Stevens, 2014b). Figure 5.5, developed by Leveson (2002), shows the relationship between the actual system and the mental models of the designer and the operator.

The designer's model is an idealization of the system before it is developed and evolves until construction is finished. Eventually, there may be significant differences between the designer's model and the actual system. The designer's model will form a basis for developing operator work instructions and training (Leveson, 2002). As shown in Figure 5.5, the operator's model is based on both the designed normative instructions and on its own experiences with the system. As discussed before, in reality, envisioned practices may differ from established practices when operators are trying to optimize their work efficiency. In addition to this, the system itself may change over time and the operator's model should be adjusted accordingly, this can only be done by working closely with the system and experiencing these changes. Often, with manual procedures that are under time pressure the limits of acceptable behavior, incorporated in the

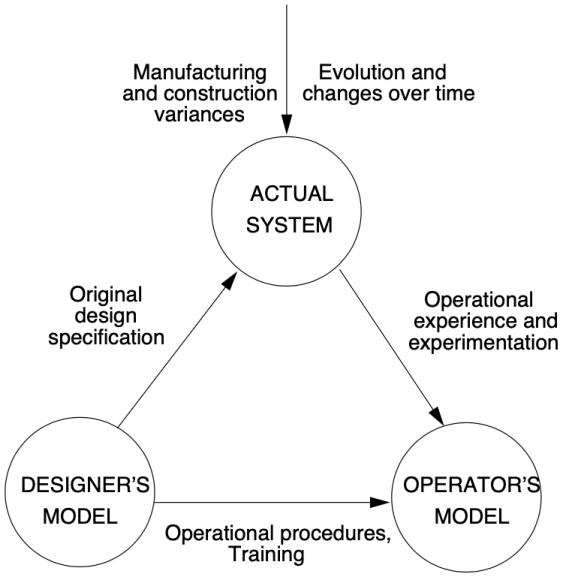


Figure L.1.: The relationship between mental models from Leveson (2002).

operator's model, can be only known from crossing the line once in a while. The value of the operator lies in the ability to adapt its model through the experiences with the system. However, often when the operator's model and its established practices are incorrect they can get the blame for flawed decision-making although this deviation may be reasonable given the provided information at that time.

As components within a system interact so does a decision maker who may depend on the activities of other decision makers (Rasmussen, 1997). This may result in accidents as it can be hard for an individual decision-maker to determine the influence of daily operational decision-making on the overall system. Decision makers do their best to use local judgment criteria and their own work environment to make the best decision, but still, this can potentially harm the system as a whole. In order to cope with this, Rasmussen (1997) states that in order to design more effective accident models a shift is needed from seeing decisions and actions as an isolated phenomenon to focusing on the mechanisms and drivers that shape human behavior and taking the context in which the behavior takes place into account. Shifting this perspective will shine a light on understanding behavior and the dynamic context, Leveson (2002) adds that this approach must include the objectives of the individual decision maker in the actual context, the boundaries of acceptable performance, the need for experimentation, and subjective criteria guiding change. This approach will view the control of human performance by identifying the boundaries of safe performance rather than focusing on deviating from normative instructions and rules set up by designers.

Lesson 4: ***Operator behavior is a product of the environment in which it occurs. To reduce operator error we must change the environment in which the operator works.***
The fourth lesson is concerned with human, or operator, error and suggests that operational behavior should be viewed together with the environment it takes place in, in order to reduce accidents. Traditionally, in attempting to reduce accidents safe behavior is rewarded and unsafe behavior punished. However, behavior and decisions are a product of the environment

which is often ever-changing. Operators that need to make decisions interact and depend on other decision-makers, in order to guide safe behavior the dynamic context should be taken into account. In addition to this, it should be acknowledged that operators develop established practices due to the optimization towards cost-effectiveness often laid down by management. Punitive measures, in case of an accident, would not serve justice when only taking into account the deviation from the normative rules set up by the designers of the system. Approaches that could handle this can be the alignment of mental models between different parties or identifying the boundaries of safe behavior.

Software Errors

Physical machinery has become less important with the introduction of software and digital automation. Although this may seem like an improvement, software control is one of the common factors involved in system accidents. The software requires enormous changes to the requirements of existing accident models and asks for new systems engineering techniques (Leveson, 2002). The usage of computers has introduced new types of accidents as well as increased the difficulty of tracing back accidents and preventing them. For the design of software systems, there is often an additional step needed to communicate the requirements from the designer to the software engineer, this is the source of the most serious errors within software applications today (Leveson, 2002). These errors and miscommunication can be traced back to flawed, or incomplete, requirements and not development mistakes such as coding errors. Such mistakes take place when the software engineer executes his perception which is different from the designer's perception.

Although, the alignment of mental models and requirements is also relevant for physical systems there are also software-specific issues that can cause harm. Software failure is not similar to physical system failure as software failure accidents stem from the operation of the software and not the lack of operations or dysfunctional ones, such as hardware issues, as do arise in physical systems (Leveson, 2002). The most significant problem with the software, however, is the curse of flexibility. Due to the computational power of computers the physical constraints, previously apparent with physical machinery, are eliminated. Physical constraints limited complexity and enforced disciplined design, construction, and modification of the technical systems (Leveson, 2002). Software being able to go beyond human intellectual limits has caused many unmanageable system accidents as those systems are often interactively complex and tightly coupled resulting in unsafe interactions that are undetectable by humans during the development and testing phases (Leveson, 2002). To overcome the issues and unsafe behavior of software, Leveson (2002) advocates for the design of tools usable for experts in the system design in which the software will operate to specify and evaluate the behavior. However, Leveson (2002) states that with the emergent properties exhibited by software due to the complex interactions it is hard to apply safety approaches and establish accident models.

Lesson 5: ***Highly reliable software is not necessarily safe. Increasing software reliability or reducing implementation errors will have a limited impact on safety.***
The fifth lesson is concerned with the fact that software failure is not similar to system failure. This is because software issues stem from the operations and the environment around the software and accidents do not arise from software actually failing itself, such as in physical systems. Therefore, it is important to realize that increasing the reliability and reducing the implementation errors will not enhance safety and the interactions and environment of the software should be incorporated in the accident models.

Adaptation

The final remark on the model of the accident is about the ability to adapt to change, especially

when social systems and humans are involved. Within organizations, there is a continuous change often charged by optimizing towards cost-effectiveness and productivity. This will result in the degeneration of safety defense systems, such as accident models, when adaptation is not built in. Leveson (2002) argues that adaptation can be predictable and controllable as it is an optimization process depending on search strategies. Rasmussen (1997) states that accidents reflect a systematic migration or organizational behavior to the boundaries of safe behavior and are fueled by a competitive environment. Adaptation should be ingrained in the safety culture of an organization by examining the behavior-shaping factors in the environment (Leveson, 2002). This approach requires organizations to look beyond events and consider the processes involved in accidents, implying that there is no deterministic root cause of accidents (Leveson, 2002). Rasmussen (1997) adds to this that causal models do not successfully incorporate the organizational and social factors in the highly adaptive socio-technical systems of today (Leveson, 2002). Finally, Leveson (2002) concludes that accident causation must be viewed as a process while taking into account the entire socio-technical system including the institutional, social, and technical environment.

Lesson 6: ***Systems will tend to migrate towards states of higher risk. Such migration is predictable and can be prevented by appropriate system design or detected during operations using leading indicators of increasing risk.***
The sixth lesson focuses on adaptation and tries to incorporate this into the safety perspective. Future accidents can be accounted for as both organizations and humans continuously change and adapt to their environment, this change can be predictable and controllable. This opposes the traditional view of accidents as assemblies of simultaneous random events. Over time safety defense systems degenerate and this should be incorporated into the system design and operational design.

M. AI System Safety Implications and Strategies

These system safety strategies for AI systems, linked to Leveson Lessons, from (Dobbe, 2022)

	Leveson Lesson	AI System Safety Implication	Example System Safety Strategy
1	Component reliability is insufficient for safety	Identify and eliminate hazards at system level	System hazard-informed system design and safety control structure
2	Causal event models cannot capture system complexity	Understand safety through socio-technical constraints	System-theoretic accident models: integrating safety constraints, the process model and the safety control structure
3	Probabilistic methods don't provide safety guarantees	Capture safety conditions and requirements in a system-theoretic way	Process model: AI system goals, actions, observation and model of controlled process and automation
4	Operator error is a product of the environment	Align mental models across design, operation and affected stakeholders	Leveson's design principles for shared human-AI controller design: redundancy, incremental control and error tolerance
5	Reliable software is not necessarily safe	Include (AI) software and its organizational dependencies in hazard analysis	System-theoretic process analysis
6	Systems migrate to states of higher risk	Ensure operational safety	Feedback mechanisms (audits, investigations and reporting systems)
7	Blame is the enemy of safety	Build an organization and culture that is open to understanding and learning	Just Culture

Figure M.1.: Overview of Leveson lessons and implications for AI with the suggested system safety strategies. Taken from Dobbe (2022).

N. Recommendations Elaborated

N.1. Embed Explainability in the company culture

First and foremost, explainability should be approached as a system property that must be incorporated and controlled throughout all layers of the organization (i.e. system), starting with the culture and company priorities. Management must be aware that designing explainable systems will help the future organization and will pay itself off, especially as most systems are becoming increasingly complex over time. It is important to realize that there will be future explainability requirements, as for example within the newly proposed EU AI act - the first law on AI by a regulator. Often management is expressing concerns for explainability, but to make change the concerns should be translated into true priorities by allocating resources. Therefore, the following three practices should be established within organizations using ML systems: develop policies and procedures on how to approach explainability, detail the explainability goals and actions, allocate resources, and assign responsibility and authority.

Develop policies and procedures
Companies mostly strive for cost-optimizations, so to realize the desire for establishing explainability practices this should be translated into the company vision. The basic underpinning of tackling emergent properties is that constraints established at a higher level, define behavior at a lower level (Bouwmans, 2020). Decisions and behavior at the lowest level of, operations, are defined in the decisions above and even outside the company at the regulatory level. Therefore, companies must specify their view on explainability and translate this into company policies and operating standards.

Detail the explainability goals and activities
The eventual goal of explainability should be determined, what information do the involved stakeholders need. This is often highly dependent on the sector and industry the company is operating in. The need for explainability is influenced by latent dimensions, discussed in Subsection 4.2.3, such as the domain and severity of incompleteness. For example, the biotechnology sector might desire causal attribution to discover new patterns in molecules and a financial institution issuing loans might only desire to justify the verdict based on the most important factors. The overall reasons for explainability, discussed in 4.2, must be established company-wide. This might differ per department and process within larger companies, however, the goal is to think about the main priority of why explainability must be included. Does the company need explainability to justify, control, improve, or discover? Companies or departments often do not know what they want, therefore, it is crucial to think about this so that activities can be adapted accordingly.

Allocate resources and assign responsibility and authority
Explainability may not always be high enough on the priority list of model developers (or management), as some may prefer to spend their efforts on improving the performance and creation

of innovative models which can optimize the company's efficiency and decrease costs. Explainability is not trivial and is a hard problem to solve. Therefore, the company must reserve the capacity for employees to work on these issues and approach explainability as a must-have.

By assigning direct responsibility and authority, there will be increased efforts due to liability. Next to this, it can increase the participation of other team members to think along. Also, including a variety of employees within such 'explainability teams' can provide significant design contributions and can potentially bring new insights. For example, when including the operator within explainability practices will provide unseen hands-on experience, and additionally, such participation can lead to a less repetitive job for the operator while making the system more effective. Also, a clear allocation of responsibility will increase the coordination among different controllers (i.e. employees). Explainability must be approached system-wide which can result in overlapping boundaries of responsibility which can beg the question of who is actually in charge, therefore responsibility must be assigned.

N.2. Create an explainability development plan

Building upon the existing policies, procedures, goals, resources, and responsible actors for the explainability practices a structured approach should be developed to translate what is necessary to approaches that are fitting the purpose and determine how to implement this. This can be done by creating an explainability development plan, this will consist of three main elements: align mental models with the involved stakeholders, perform research on suitable methods for explainability, and research the technical limitations and hazards of these methods.

Align the mental models of the involved stakeholders
Taking software engineering as an example, research from Leveson and Weiss (2009) shows that the vast majority of software-related accidents can be traced back to flawed requirements. In order to formulate requirements, it is important that the stakeholders know what they actually want. Aligning the mental models will allow stakeholders to think about this. The empirical research within TM has shown that often non-technical stakeholders have difficulties providing requirements or probing and assessing those of more technical stakeholders such as the model developers. The alignment of mental models can help to close the gap between eventual normative procedures and established procedures by being aware of the environmental factors and technical limitations of the system. This will bring all stakeholders closer together and manage expectations on the ability and limitations of existing explainability approaches.

Perform research on suitable methods for explainability and limitations
Next, the suitable existing explainability methods which can potentially be applied in the system must be researched. For example, global and local explainability approaches should be researched by their ability to handle the type of data, type of model, and whether it can potentially fulfill the requirements of the system's use case.

The potential methods should be evaluated on their limitations, ensuring that the limitations are acknowledged and known, and measured against properties of good explanations and explanation techniques described in Subsection 4.2.4 and Subsection 4.2.5. Next, the potential hazards coming with these limitations should be mapped. And lastly, the potential hazards that can come up by using the explainability method should be determined from a socio-technical perspective. Examples of such hazards can be the over-reliance of operators on certain local

explainability methods, the hazard of deviation between normative work instructions and procedures described in operational documentation and established practice, or that model documentation must be updated accordingly after model retraining.

N.3. Operationalize explainability using a user-centered approach

Now that the policies and operating standards are established, there is a clear allocation of responsibility, the mental models are aligned, and the possible techniques and limitations are known the established method can be used. The method for a user-centered operationalization of explainability is pictured in Figure 7.2 and elaborately discussed in Chapter **??**.

N.4. Install structured communication channels

In order to measure the satisfaction of explainability constraints, to align mental models, and to provide feedback communication channels need to be established. Next to this, the empirical study has shown that providing uniform elements within documentation can help stakeholders understand concepts and minimize keyperson risk.

Foster and design communication channels throughout the entire organization
One of the primary roles of system engineering is establishing and implementing technical communication channels. Once again similarities can be noticed from system safety accident analysis. Multiple accident reports in software-related accidents show that often the problem leading to losses is already visible, but there were no communication channels established for transferring the information to those who can solve this, or, that the measuring channel was ineffective or unused (Leveson, 2001). Learning from this, there should be structured communication channels that will serve to set communication throughout the organization and not merely in control structures or between neighboring hierarchical levels. The information gained at the lowest level of operations can provide valuable insights for the development of new techniques, an adaptation of work instructions, and operating procedures, and might even influence company policies and standards. This will avoid factors such as time and productivity pressure coming in the way of communicating. The structured channels will force to get the stakeholders around the table to evaluate current practices and allow for a culture where there is time created to talk about potential issues.

Provide uniform communication means to enhance understanding
Next to this, the more formalized means of communication, such as the measuring channel in Figure 5.1, should entail uniform elements across all designed models (if possible). When the model checklists do include similar technical elements to be investigated or audited it will increase the shared understanding, and anticipate employee turnover or changes over time when the checklists or model documentation have a shared structure (I9). Also, this will make sure that people get familiar with the terminology and requirements and will avoid that knowledge or the design rationale is being lost with the departure of the creators. This will also enable auditors or regulators to compare models or operations to adapt to new changes in model specific elements more easily.

N.5. Avoid complexity and re-think the actual objective

Taking a step back and rethinking how explainability has even become an issue at all leads to the last recommendation. The unconstrained technical ability of ML to add features or develop combinations of models together with the desire to optimize performance within organizations can result in the design trap of creeping featurism. This is a concept describing the systematic tendency to add or expand a product with additional features making the system become more complex (Winograd and Woods, 1997). This is one of the main concerns raised by practitioners. By striving for performance optimization of the models, such as TM models, they become more complex and use more and more features resulting in greater complexity of both the model and the system as a whole. This results in an overall system where it is harder to test, provide explainability, audit, review, and maintain while costs are rising (Leveson and Weiss, 2009). Model developers need to refrain from compelexity and must make hard decisions on the functionality of models while taking into account the effectiveness, explainability, and maintenance costs. Avoiding complexity results in designing for the exact model objectives, stakeholder requirements, and keeping the explainability reasons of the audience in mind.

The example of TM shows that due to the regulatory pressure models have been developed with the priority to catch as many potentially suspicious clients as possible, however, now that costs, capacity, and explainability become an issue this calls for a change. Designing more simple models focused on effective specific SIRA scenario detection will result in lower maintenance and higher explainability (I5). A suggestion could be to introduce modular models, where all models have the basic set of features and different models can have SIRA risk specific add-on features (I2, I5, I9). This will enable the ability to evaluate the intended model goal and allow for a comparison between models (I5). More importantly, it will allow for more profound explainability practice, as current techniques such as SHAP are having difficulties handling complexity (I3, I6). But, limiting the complexity can have a negative influence on performance, a balance should be sought between effectiveness, overall performance, and intrinsic explainability in order to make an informed decision.

O. Interview Codes and Stakeholder roles

AI System Role	TM Stakeholder
Creator-Owner:	Risk Owner
Creator-Implementer:	Data scientist
Operator:	TM analyst
Executor:	TM analyst
Decision Subject:	Client (Business or Natural Person)
Data Subject:	Clients (Businesses or Natural Persons)
Examiner:	(Internal regulators) 2nd line of defence: Model Validation, Legal, Compliance. 3rd line of defense: Audit. (External regulators) DNB, AP

Table O.1.: Stakeholder roles in transaction monitoring

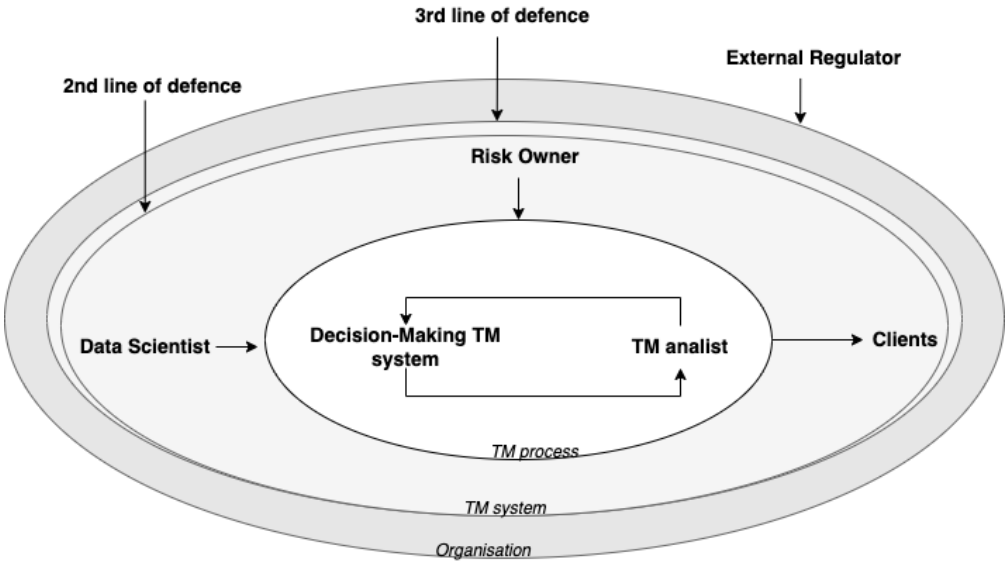


Figure O.1.: Stakeholders within the transaction monitoring process

Code	Sub-category	Amount of references
Current Explainability Approaches	Institutional Social Technical	15 5 9
Reasons for Explainability	Control Justify Improve Discover	I4 9 7 9
Hazards of Explainability Methods and Usage	Institutional Social Technical	15 34 25
Improvements of Explainability Approaches	Instituional Social Technical	37 5 18

Table O.2.: Interview codifications and references

P. Toy Case: method demonstration and evaluation

The designed case for the focus group is pictured in Figure P.1 and the illustrative control structure is presented in Figure P.2.

P.1. Toy Case

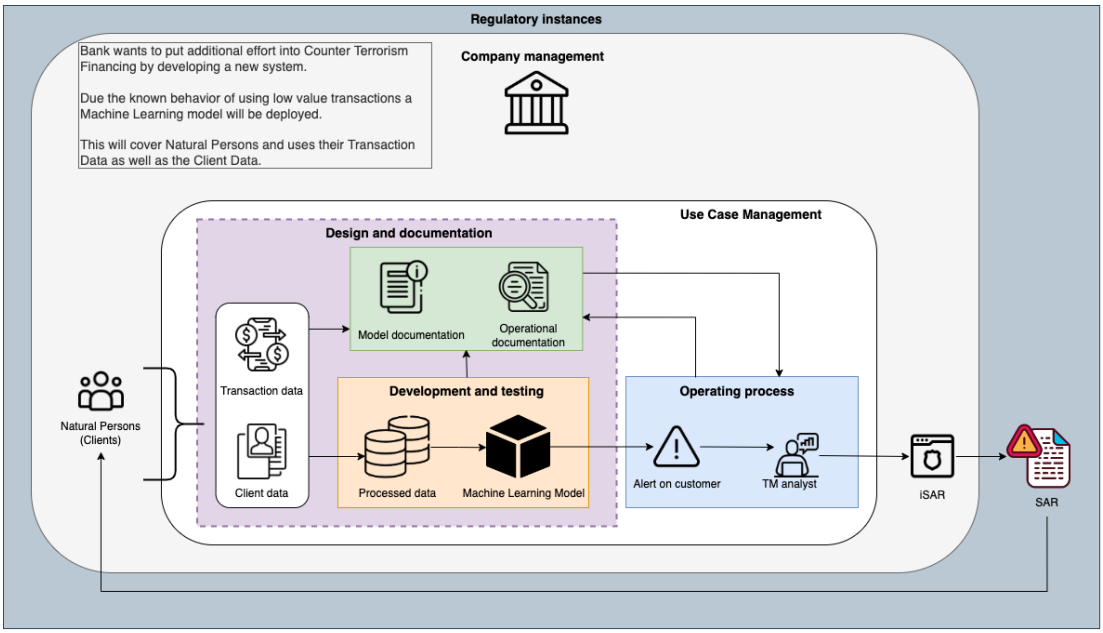


Figure P.1.: Toy Case provided to the local practice focus group

P.2. Evaluation

For the evaluation two questions are asked to the attendees of the focus group. The questions, together with the answers are formulated below. Note that not all attendees answered any question or both, the attendees are remained anonymous.

Question 1: How did the method help you in establishing explainability?

Answer to Q1: Useful exposition of all the different components, invites a new way of thinking about explainability that enables actions. It would enable design processes both in general and per use case.

Answer to Q1: Exhaustive in terms of fully defining the explainability constraints. I think that using the method the model builder examines the contraints with more structure. I think the workshop made me think of an angle I did not think of before.

Answer to Q1: It is very useful to divide explainability into three components, this provided a good structure to think about it and talk about it. The method resulted in a lot of new insights. I think the workshop was a great experiment!

Answer to Q1: As a process manager it's useful to have this method to refer to when engaing with stakeholders on explainability requirements. The constraints (downwards) and feedback (upwards) setup is easy to understand and helpful.

Answer to Q1: To look at explainability from a broader perspective.

Answer to Q1: Very concrete to create the method for explainability. Brings up a lot of questions and dilemma's.

Answer to Q1: To think more thoroughly about explainability and not only as SHAP output. There is a larger process than only the data science teams.

Question 2: What could be improved?

Answer to Q2: Perhaps additional guidance could be provided how to instantiate the control structure.

Answer to Q2: Make the stakeholder in control levels explicit.

Answer to Q2: Maybe also consider the costs within the method, explainability methods can be computational expensive.

Answer to Q2: Who to ask the questions? As a data scientist I can only reason so much about some aspects of explainability.

Answer to Q2: Examples within the different layers in the hierarchical control structure.

Answer to Q2: It was a complex framework to understand it all at once. (Requiring to understand the method a bit in depth in order to answer). Perhaps the questions could be 'dumbed down' to help answer them for applications in practice.

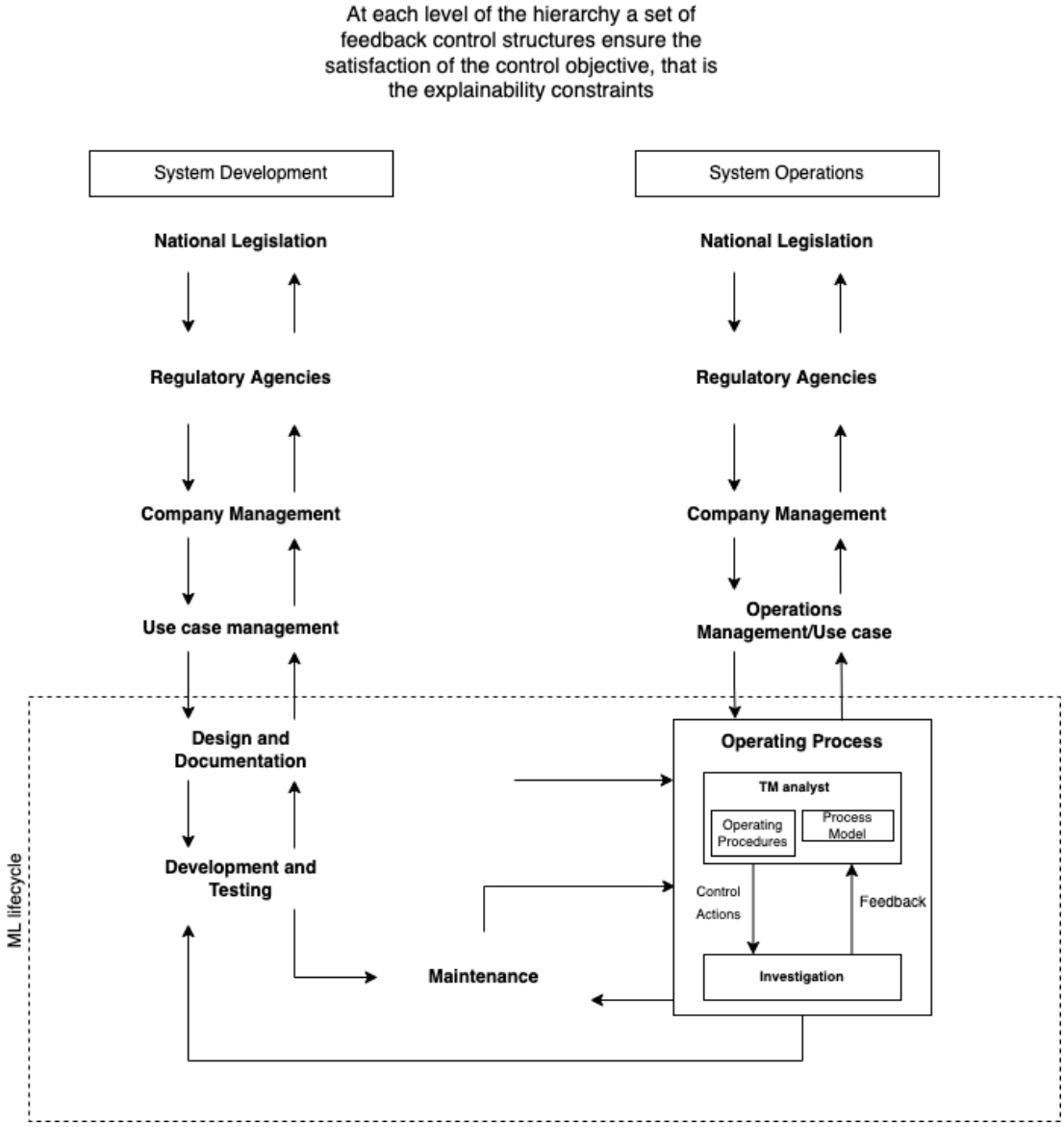


Figure P.2.: Control structure provided to the local practice focus group

Bibliography

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., and Hoos, H. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(08):18–28.

Allen, G. (2020). Understanding ai technology. Report, Joint Artificial Intelligence Center (JAIC) The Pentagon United States.

Alon-Barkat, S. and Busuioc, M. (2022). Human-ai interactions in public sector decision-making:'automation bias' and 'selective adherence'to algorithmic advice. *Journal of Public Administration Research and Theory*.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.

Ashby, W. R. (1957). An introduction to cybernetics.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Benk, M., Tolmeijer, S., von Wangenheim, F., and Ferrario, A. (2022). The value of measuring trust in ai-a socio-technical system perspective. *arXiv preprint arXiv:2204.13480*.

Bouwmans, I. (2020). Designing in socio-technical systems. page CH7.

Bryson, J. J. (2019). The past decade and future of ai's impact on society. *Towards a new enlightenment*, pages 150–185.

Buijsman, S. and Veluwenkamp, H. (2022). Spotting when algorithms are wrong. *Minds and Machines*, pages 1–22.

Campbell, M., Hoane, A. J., and Hsu, F.-h. (2002). Deep blue. *Artificial Intelligence*, 134(1):57–83.

Castro, P., Rodrigues, J. P., and Teixeira, J. G. (2020). Understanding fintech ecosystem evolution through service innovation and socio-technical system perspective. In *International Conference on Exploring Services Science*, pages 187–201. Springer.

Checkland, P. (1981). Systems thinking, systems practice.

Checkland, P. (2012). Four conditions for serious systems thinking and action. *Systems Research and Behavioral Science*, 29(5):465–469.

Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., and Lam, K. S. (2018). Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, 57(2):245–285.

de Bruijn, H., Warnier, M., and Janssen, M. (2021). The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, page 101666.

De Nederlandsche Bank (2020). Leidraad wwft en sw.

de Souza Nascimento, E., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., and Conte, T. (2019). Understanding development process of machine learning systems: Challenges and solutions. In *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–6. IEEE.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Dobbe, R. (2022). System safety and artificial intelligence. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1584–1584.

Dobbe, R., Gilbert, T. K., and Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Duan, Y., Edwards, J. S., and Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management*, 48:63–71.

Edwards, J. S., Duan, Y., and Robins, P. C. (2000). An analysis of expert systems for business decision making at different levels and in different roles. *European Journal of Information Systems*, 9(1):36–46.

European Parliament and the Council (2015a). Directive (eu) 2015/849 of the european parliament and of the council of 20 may 2015 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, amending regulation (eu) no 648/2012 of the european parliament and of the council, and repealing directive 2005/60/ec of the european parliament and of the council and commission directive 2006/70/ec.

European Parliament and the Council (2015b). Regulation (eu) 2015/847 of the european parliament and of the council of 20 may 2015 on information accompanying transfers of funds and repealing regulation (ec) no 1781/2006.

European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal L110*, 59:1–88.

Europol (2022a). About europol. https://www.europol.europa.eu/about-europol. Accessed: 04-05-2022.

Europol (2022b). Economic crime. https://www.europol.europa.eu/crime-areas-and-statistics/crime-areas/economic-crime. Accessed: 04-05-2022.

Europol (2022c). European financial and economic crime. https://www.europol.europa.eu/about-europol/european-financial-and-economic-crime-centre-efecc. Accessed: 05-05-2022.

FATF (Oktober, 2014). Guidance for a risk-based approach: The banking sector. Report.

Gao, S. and Xu, D. (2009). Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering. *Expert Systems with Applications*, 36(2):1493–1504.

Gentner, D. and Stevens, A. L. (2014a). *Mental models*. Psychology Press.

Gentner, D. and Stevens, A. L. (2014b). *Mental models*. Psychology Press.

Gianfagna, L. and Di Cecco, A. (2021). *Explainable AI with Python*. Springer.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science Robotics*, 4(37).

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hayes-Roth, F. (1985). Rule-based systems. *Communications of the ACM*, 28(9):921–932.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4.

Ilkou, E. and Koutraki, M. (2020). Symbolic vs sub-symbolic ai methods: Friends or enemies? In *CIKM (Workshops)*.

Istrefi, K. and Piloiu, A. (2020). Public opinion on central banks when economic policy is uncertain. *Revue d'économie politique*, 130(2):283–306.

Jalali, S. and Wohlin, C. (2012). Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the 2012 ACM-IEEE international symposium on empirical software engineering and measurement*, pages 29–38. IEEE.

Johannesson, P. and Perjons, E. (2014). *An introduction to design science*. Springer.

Johnson, W. G. (1980). *MORT safety assurance systems*, volume 4. Marcel Dekker Incorporated.

Kaya, H. (2022). Human centered machine learning lecture slides - taxonomy of xai methods.

Kile, F. (2013). Artificial intelligence and society: a furtive transformation. *AI & society*, 28(1):107–115.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kuiper, O., Berg, M. v. d., Burgt, J. v. d., and Leijnen, S. (2021). Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities. pages 105–119.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473.

Leveson, N. (1995). Safeware: System safety and computers. addison wesley, reading.

Leveson, N. (2001). Systemic factors in software-related spacecraft accidents. page 4763.

Leveson, N. (2003). White paper on approaches to safety engineering. *Disponible en ligne sur le site de l'auteur (sunnyday. mit. edu/caib/concepts. pdf)*.

Leveson, N. G. (2002). System safety engineering: Back to the future. *Massachusetts Institute of Technology*.

Leveson, N. G. (2011). *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

Leveson, N. G. and Stephanopoulos, G. (2013). A system-theoretic, control-inspired view and approach to process safety.

Leveson, N. G. and Weiss, K. A. (2009). Software system safety. pages 475–505.

Liao, Q. V. and Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266.

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

Luft, J. and Ingham, H. (1961). The johari window. *Human relations training news*, 5(1):6–7.

Lughofer, E., Richter, R., Neissl, U., Heidl, W., Eitzinger, C., and Radauer, T. (2017). Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior. *Information Sciences*, 420:16–36.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

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

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. pages 220–229.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mäkinen, S., Skogström, H., Laaksonen, E., and Mikkonen, T. (2021). Who needs mlops: What data scientists seek to accomplish and how can mlops help? In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 109–112. IEEE.

National Transportation Safety Board (2010). Introduction of glass cockpit avionics into light aircraft. *Safety Study*.

Nations, U. (2000). United nations, vienna convention on the law of treaties, 23 may 1969, united nations, treaty series, vol. 1155, p. 331, available at: https://www.refworld.org/docid/3ae6b3a10.html [accessed 12 november 2022].

Nesvijevskaia, A., Ouillade, S., Guilmin, P., and Zucker, J.-D. (2021). The accuracy versus interpretability trade-off in fraud detection model. *Data & Policy*, 3.

Nicholls, J., Kuppa, A., and Le-Khac, N.-A. (2021). Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape. *IEEE Access*.

Offermann, P., Blom, S., Schönherr, M., and Bub, U. (2010). Artifact types in information systems design science–a literature review. In *International Conference on Design Science Research in Information Systems*, pages 77–92. Springer.

Ostheimer, J., Chowdhury, S., and Iqbal, S. (2021). An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles. *Technology in Society*, 66:101647.

Padovan, P. H., Martins, C. M., and Reed, C. (2022). Black is the new orange: how to determine ai liability. *Artificial Intelligence and Law*, pages 1–35.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., and Turini, F. (2019). Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9780–9784.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety science*, 27(2-3):183–213.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rittel, H. W. and Webber, M. M. (1974). Wicked problems. *Man-made Futures*, 26(1):272–280.

Robnik-Šikonja, M. and Bohanec, M. (2018). *Perturbation-based explanations of prediction models*, pages 159–175. Springer.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85.

Rutte, M., Kaag, S., Hoekstra, W., and Seger, G. (2021). Omzien naar elkaar, vooruitkijken naar de toekomst. Report.

Sarkara, A. (2022). Is explainable ai a race against model complexity?

Sartori, L. and Theodorou, A. (2022). A sociotechnical perspective for the future of ai: narratives, inequalities, and human control. *Ethics and Information Technology*, 24(1):1–11.

Schlossberger, O. (2015). *Anti-Money Laundering*. EU Press.

Schott, P. A. (2006). *Reference guide to anti-money laundering and combating the financing of terrorism*. World Bank Publications.

Selbst, A. D. and Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87:1085.

Sharman, J. C. and Chaikin, D. (2009). Corruption and anti-money-laundering systems: putting a luxury good to work. *Governance*, 22(1):27–45.

Silva, P. G. (2019). Recent developments in eu legislation on anti-money laundering and terrorist financing. *New Journal of European Criminal Law*, 10(1):57–67.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Sun, T. Q. and Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, 36(2):368–383.

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.

United Nations (1969). United nations, vienna convention on the law of treaties, 23 may 1969, united nations, treaty series, vol. 1155, p. 331, available at: https://www.refworld.org/docid/3ae6b3a10.html [accessed 12 november 2022].

van de Poel, I. (2020). Embedding values in artificial intelligence (ai) systems. *Minds and Machines*, 30(3):385–409.

van Huffelen, A. C. (2021). 8e voortgangrapportage kinderopvangtoeslag. Government document, Belastingdienst - Toeslagen en Douane.

Verma, A. K., Ajit, S., and Karanki, D. R. (2010). *Reliability and safety engineering*, volume 43. Springer.

Wagner, B. (2019). Liable, but not in control? ensuring meaningful human agency in automated decision-making systems. *Policy Internet*, 11(1):104–122.

Wee, B. V. and Banister, D. (2016). How to write a literature review paper? *Transport Reviews*, 36(2):278–288.

Weinberg, G. M. (2001). *An introduction to general systems thinking (silver anniversary ed.)*. Dorset House Publishing Co., Inc.

Wilhelm, W. K. (2004). The fraud management lifecycle theory: A holistic approach to fraud management. *Journal of economic crime management*, 2(2):1–38.

Winograd, T. and Woods, D. (1997). The challenge of human-centered design. *Human-centered systems: information, interactivity, and intelligence*.

World Economic Forum (2022). Global coalition to fight financial crime. https://www.weforum.org/projects/coalition-to-fight-financial-crime. Accessed: 04-05-2022.

Young, W. and Leveson, N. G. (2014). An integrated approach to safety and security based on systems theory. *Communications of the ACM*, 57(2):31–35.

Zhu, X., Ao, X., Qin, Z., Chang, Y., Liu, Y., He, Q., and Li, J. (2021). Intelligent financial fraud detection practices in post-pandemic era. *The Innovation*, 2(4):100176.