Universiteit Leiden
The Netherlands

Informatica & Economie

Towards a Practical Solution for Ensuring Traceability of Automated Outcomes in the Dutch Insurance Industry.

Kas Kansak Bekker

Supervisors:
Prof. Joost Visser, Robert Akkerman & Walter Mosterd

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

Abstract

In this thesis, we researched the most practical solution to ensure traceability of automated outcomes in the Dutch insurance industry. Following the Design Science Research methodology, exploratory interviews and a literature study were conducted. The literature study provided a definition of traceability, the urgency of traceability and discusses various related work regarding requirements for traceability. Together with the SIVI foundation and industry experts, we have identified a base set of compliance and management focal points fitted specifically to the Dutch Insurance Industry. It can be concluded, taking into account the current state and use of automated outcomes, that at the moment of writing, the most practical solution would be our designed artifact: a litepaper. This litepaper is a businesses document, mapping out various requirements for traceability and discussing practical focal points for achieving these requirements. Finally, confirmatory interviews were conducted validating the litepaper among the interviewed industry experts. Resulting in the conclusion that the litepaper should give a give a good overview of what to account for when implementing traceability.

Contents

1 Introduction

The use of automated applications, e.g. Artificial Intelligence (AI) and Machine Learning (ML) techniques are rising in the Dutch insurance industry. Decisions and predictions based on using automated decision making can be highly valuable. The Autoriteit Financiële Markten & De Nederlandsche Bank (AFM & DNB, 2019) recognize these developments and point out that the use of these new developments not only offer new opportunities, but also uncertainties and risks. According tot the AFM and DNB, adequate attention and awareness of these uncertainties and risks is needed in order to enable the deployment of responsible AI, compliant with the requirements regarding reliable and controlled operations, product development and the legislation surrounding the duty of care.

With this increasing usage of AI including its associated uncertainties and risks, Leijnen et al. (2022) argue that there is a call to remain understandable and transparent. Certain decisions, predictions or outcomes become more difficult to explain or identify how they have been established. How does an AI solution come to a certain outcome or decision? How do you explain such an outcome, considering the huge amount of data used? This research area is called explainable AI, or in short XAI. Leijnen et al. (2022) gives the following definition of XAI:

> “XAI is aimed to provide a solution to the ‘black box’ problem in AI. That is, an AI solution utilizes data and produces an outcome. However, in this process there is generally no output that explains how or why the outcome is reached based on the data. Especially in the case of AI techniques such as deep neural networks, the process from input to output is virtually impossible to interpret even with knowledge of the inner workings, weights, and biases of the system. XAI explains why or how the AI solution arrived at a specific decision, “

In our research, we will focus on a part of the research area of XAI: traceability. According to Leijnen et al. (2022), traceability of an automated outcome entails it can be determined which data is used, by what process the outcome is generated and most importantly the documentation of the processes (see more in 4.2.1). Leijnen et al. (2022) point out the importance of traceability as follows:

> “It is especially important in relation to auditability, the trait of being capable of being audited. Traceability is especially important in medical, financial, and other domains with strong legislation and potential ethical risks. For other domains traceability might be less important, e.g. an AI solution that detects unripe fruit to remove from a conveyor belt which does not directly impact humans. ‘

Considering its importance, the urgency of traceability specifically emerges when things go wrong. For example, in 2022 a tool from Hypotheken Data Netwerk (HDN) calculated the test income of a consumer wrong for a mortgage application for a vast period of time, resulting in a lot of work of correcting these applications retroactively (Ettema, 2021b).

To mitigate the risk surrounding the use of automated decision making, there are two regulatory organizations for the financial sector called the Financial Markets Authority (AFM) and the Dutch Central Bank (DNB). An exploration on the use of AI in the Dutch Insurance Industry published by the AFM and DNB reveals several concerns. AFM and DNB (2019) stipulate that the application of AI techniques in business processes is still in development. Insurers should pay attention to various technical aspects of AI models when using AI. According to the AFM and DNB, a combination of organisation-wide knowledge about AI and internally developed policies on the use of AI is important to deploy AI responsibly.

In collaboration with the SIVI foundation (see more in section 5.1), we attempt to contribute to extending the knowledge of traceable outcomes and its urgency in the Dutch Insurance Industry. Here, we will focus on a practical solution to offer stakeholders hands on support based on already existing research of traceability. To break down the problem statement into a single question, we formulated the following research question:

- *"What will be the most practical solution to ensure the traceability of automated outcomes in the insurance industry?"*

To support this question, we will use the Design Science Research framework and conduct various interviews and execute a literature study. Hereafter, the interviews and literature study will support the creation of practical solution. Finally, this practical solution will be the creation of a "lite paper" containing the investigated information and will be deployed to the interviewees.

Furthermore, we will divide our research in 3 more subareas to be investigated. First, it is important to know what the definition of traceability entails (4.2) and what the relation is towards other terminology (4.2.3) in the field of XAI. Second, we will research which traceability requirements are needed to guarantee traceability and which requirements apply to the insurance industry (4.3). Finally, we will shortly consider the existing regulations surrounding traceability (4.2.2).

1.1 The situation

In this section, we will discuss the situation around the concept of traceability. First, we will give an introduction to the current situation of the umbrella topic: XAI, which is followed by its additional difficulties and problems. Thereafter, we will dive deeper into the situation specifically for the insurance industry. By discussing a few cases, we will try to give an example of the traceability challenges in practice.

With the wave of growing awareness of the use of artificial intelligence applications and its explanations, Brundage et al. (2020) recognize that current existing regulations and norms are insufficient to ensure responsible and ethical AI development. Various instances, e.g. the European Commission, have taken steps to address and operationalize this insufficiency by publishing non-binding principles to attain "Trustworthy AI" (European Commission, 2019). However, Brundage et al. (2020) argues the following: "These steps and principles are ill-equipped to assess whether an AI developer's actions are consistent with the stated principles. There is a need to move beyond principles, to focus on mechanisms for demonstrating responsible behaviour". The report "Toward Trustworthy AI Development" dives deeper into "various steps which different stakeholders in AI development can take to make it easier to verify claims about AI development" (Brundage et al., 2020). These suggested steps attempt to solve and addresses among other things the following problem:

> “ AI systems lack traceable logs of steps taken in problem-definition, design, development, and operation, leading to a lack of accountability for subsequent claims about those systems' properties and impacts “ (Brundage et al., 2020)

With the deployment of AI in safety critical contexts, e.g. a health insurance policy, audit trails will become more important. Brundage et al. (2020) describes an audit trail as a traceable log in a systems operation. ”The recommendation of steps to meet this growth is standard setting bodies, which should work with academia and indusry to develop audit trail requirements for safety critical applications of AI systems. To make safety-critical AI systems fully auditable, software audit trails often require a base set of traceability trails to be demonstrated for qualification.” (Brundage et al., 2020)

These problems as addressed by Brundage et al. (2020) also arise in the insurance industry. In an exploration of the AFM & DNB about the use of AI in the insurance industry, 10 points of interest are given divided into three categories: i) embedding AI in an organization, ii) technical points of interest of AI and iii) AI & the consumer (AFM & DNB, 2019). One of its points of interest: ”5 - explainable outcomes”, states that an insurer should be able to identify the relationships between input parameters, the model and the outcomes. Examples given are models for pricing, acceptance or fraud detection which have direct impact on the consumer when established by automated processes. The DNB and AFM state it is important to be able to trace back the individual input parameters and to which extent these contributed to the outcome.

Not being able to identify relationships between input parameters, model and outcomes can have a lot of impact. For example: in 2003 a consumer takes out an investment insurance by a major insurance group. At Kifid, a dispute committee for consumers, the consumer complains about a few points: the costs, the absence of recovery advice and the ”leverage & equity erosion”. At the first assessment[1], Kifid adjudicates in favor of the insurance company. On appeal[2], Kifid judged something else: the insurer did not give a transparent explanation about the ”mechanism” behind the calculation of the costs. Based on a directive (Artikel 5 Richtlijn 93/13/EEG[3]), in the case of written contracts, clauses needs to be clear and understandable. In the case of doubt, the stipulation of the consumer will prevail. The result: the stipulation about the costs was ruled in favor of the consumer because the insurance group could not give nor explain factors which were decisive for the height of the costs. Eventually, it led to the insurance group refunding part of the costs back to the consumer. This case was in 2003 and gives a good example about the (financial) consequences whenever factors impacting the calculations could not be traced back or reproduced.

A more recent case is an error which occured in the IBL calculation tool (Inkomensbepaling Loondienst tool). HDN (2022) describes the IBL tool as a method to calculate the test income for employees when applying for a mortgage. It uses income data of the mortgage applicant, which again is originating from the Uitvoeringsinstituut Werknemersverzekeringen (UWV). An advisor eventually reviews the applicant's test income with the use of this online IBL calculation tool (HDN, 2022). In february 2021, HDN reported that the IBL tool had given a wrong test income for

[1] https://www.kifid.nl/fileupload/jurisprudentie/GeschillenCommissie/2016/Uitspraak_2016-261_Bindend.pdf

[2] https://www.kifid.nl/wp-content/uploads/2018/07/Cvb-2018-041.pdf

[3] https://eur-lex.europa.eu/legal-content/NL/ALL/?uri=CELEX:31993L0013

a vast period of time.

Ettema (2021b), author at InFinance magazine writes the following about the incident:

- Between 13th of august 2020 and 27th of January, a wrong test income is given from the IBL tool by customers with a lease-car.
- The bug originates from a modification in a external data source (UWV as well as other IBL-data sources which weren't known to HDN.)
- After another error the IBL tool was taken offline and HDN stated that they will improve their governance, processes and controls around the IBL tools.

According to Ettema (2021b), Martin Keegsta, Director at Romeo Financial Services already had some concerns about using external data for mortgage advice and points out that it could lead to unacceptable risks. Keegstra states that the root cause of the problems with the HDN tool is that no contractual arrangements were made regarding the sources of the data. Resulting in that changes in the chain were unnoticed and mortgages with incorrect test income were granted, which need to be reviewed again (Ettema, 2021a).

To avoid problems and its consequences shown in cases such as above, various Dutch instances also develop practical guidelines or tools to contribute to operationalizing traceability and transparency. For example SIVI (2022) developed a checklist platform for quality of non-human applications to give insights to organizations how to deal with themes such as legislation, technological risk, testing, traceability and monitoring surrounding non-human applications. Another example, het Verbond van Verzekeraars published "Ethisch Kader Data Toepassingen" (Verbond van Verzekeraars, 2015) which includes norms to give extra checks to the ethical use of automated outcomes in the insurance industry. In this research we will also try to contribute to the operatizionaliztion of ensuring traceability in practice.

1.2 Thesis overview

This thesis is structured in four parts. The first sections, related work as well as our methodology are focused on research practicalities and are meant to show which earlier relevant research is done and how this research is executed. Thereafter, we will use the section "literature study" (see section 4) to elaborate on the urgency of traceability by outlining current guidelines and legislation, discussing its definition and eventually its requirements. Thereafter, we will present our exploratory interviews, which are meant to set the needs of the Dutch Insurance Industry relating guaranteeing traceability. Finally, we will present a practical solution of ensure traceability in the results section and discuss why and how this artifact will function as practical solution.

2 Methodology

In this section, we will discuss the methods used in our research. As overall methodological framework, we followed a Design Science Research approach. Within this framework, we first conducted exploratory interviews In order to get a broader understanding of the specific domain. With the findings of these exploratory interviews, we started a literature study based on the questions to be investigated as well as the findings of the exploratory interviews. Following the approach of Design Science research, we created an artifact based on the findings of the exploratory interviews and the literature study. this artifact is a litepaper that contains our findings in the form of some recommendations.

2.1 Design Science Methodology

For this research, we will follow a design science methodology, or in other words the Design Science Research (DSR) methodology. DSR is a problem-solving methodology that bridges the gap between a problem statement in the social context and providing a solution with research from knowledge context. Most of the time, this solution is described as an artifact. These artifacts are represented by constructs, models and methods. The results of the DSR methodology include these newly designed artifacts, as well as verification and validation on how and why these designed artifacts solve the problem relevant to the social context (vom Brocke et al., 2020). The authors vom Brocke et al. (2020) have drawn their inspiration from earlier contributions to design entities (e.g. Hevner and Peffers).

2.1.1 The DSR framework

The DSR methodology is bound to a framework, which is defined by vom Brocke et al. (2020) and is described as follows:

- *The environment* (social context) defines the problem space where a challenge or problem arises. This context is composed of various actors such as people, organizations but also information which is already available such as technology, policy documents etc. vom Brocke et al. (2020) state that in this context, needs are assessed and evaluated after which these needs are positioned relative to the existing technology, applications, communication architectures and development capabilities. Together these define the "research problem" (vom Brocke et al., 2020). In our research, the environment and social context consists of all stakeholders (suppliers, insurers and supervisory bodies) in the Dutch Insurance Industry.

- *The knowledge base* (knowledge context) provides materials, such as prior research, theories or frameworks through which the artifact can be substantiated. In this research, the knowledge base will consist of a Literature Study focusing on various knowledge areas surrounding traceability (urgency, legislation, technical execution).

- *Design*. In the design stage, the need for a solution is linked with the specific "knowledge base". The Design Science Research is set out to create an innovative solution to the problem. For establishing the "needs" and identifying the linked knowledge, existing diverse research methods can be applied (e.g. interviews, surveys, literature reviews). With these needs, we will use the knowledge base to design and develop an artifact. Finally this results in an artifact which can be tested and evaluated (vom Brocke et al., 2020). In our research, we will design an artifact which will be a litepaper and will consist of around 20 pages, clarifying the current requirements and practical documentation solutions for traceability.

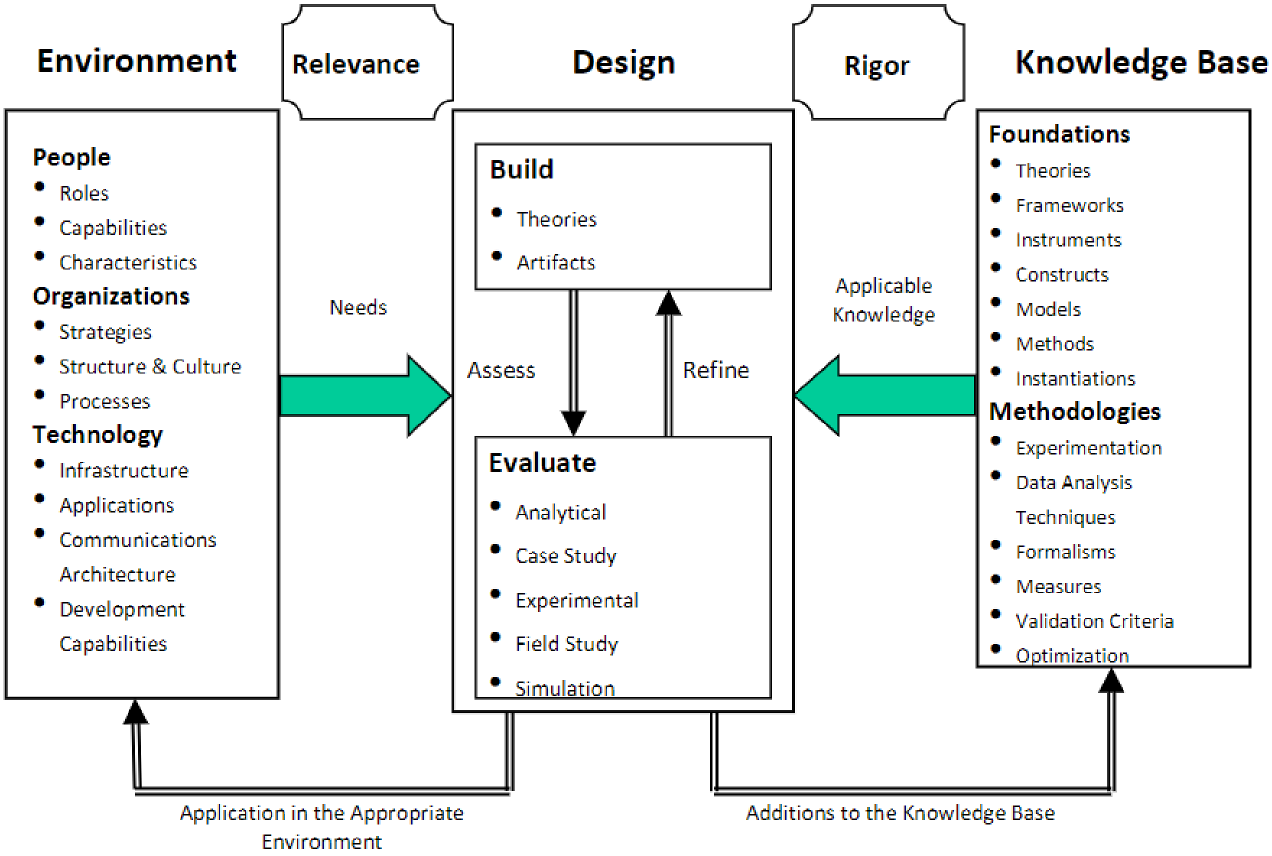


Figure 1: Design Science Research Framework (vom Brocke et al., 2020)

2.1.2 The DSR Process

The Desgin Science Research Methodology follows the following process, including six steps which are defined by vom Brocke et al. (2020) as follows:

1. *Problem Identification*, here the specific research problem as well as the value is identified. For our research, the specific research problem was driven by the need of *Stichting SIVI* to get a broader understanding about the topic of traceability. Here we identified the urgency of traceability by researching various cases as well as existing guidelines and regulations.

2. *Define the objectives for a solution*, here the objectives of a solution is inferred from the problem definition and knowledge. For our research, we inferred the objectives with exploratory interviews held by different stakeholders in the Dutch Insurance Industry.

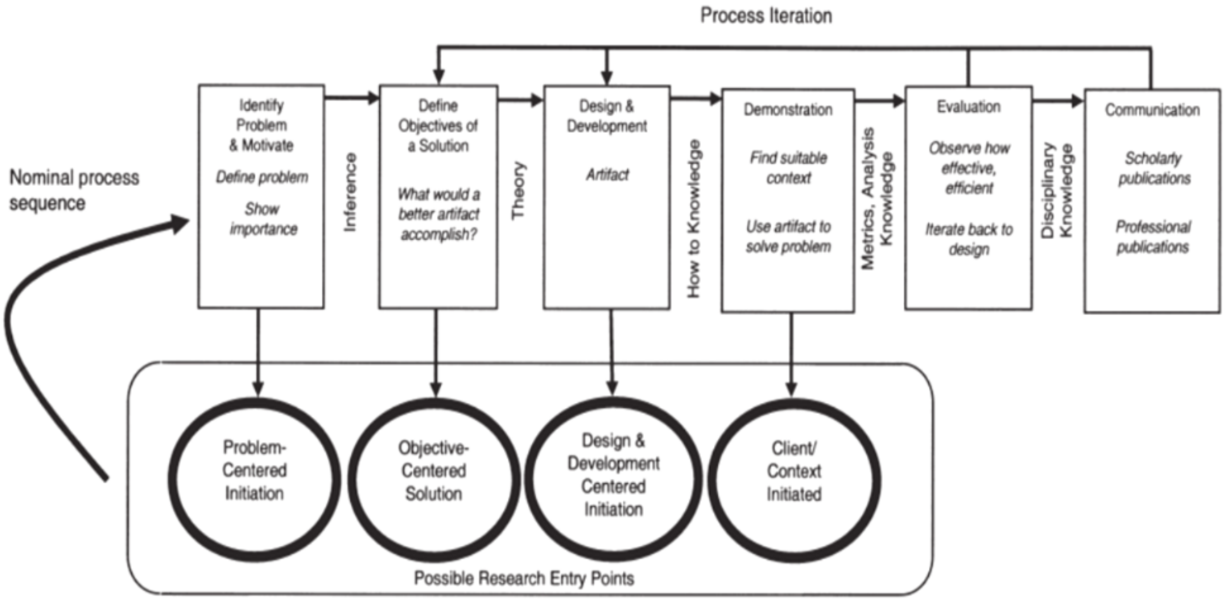

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Define problem
Show importance
Inference
Define Objectives of a Solution
What would a better artifact accomplish?
Theory
Design & Development
Artifact
How to Knowledge
Demonstration
Find suitable context
Use artifact to solve problem
Metrics, Analysis Knowledge
Evaluation
Observe how effective, efficient
Iterate back to design
Disciplinary Knowledge
Communication
Scholarly publications
Professional publications
Problem-Centered Initiation
Objective-Centered Solution
Design & Development Centered Initiation
Client/ Context Initiated
Possible Research Entry Points

Figure 2: Design Science Research Framework (vom Brocke et al., 2020)

3. *Design and Development*, in this stage an artifact is created. Because traceability is a still evolving topic, we chose for an artifact which would fit the needs as identified as well as possible, but is still open for any future changes. In our case the artifact will be a litepaper, a short (20 pages) explainable business report, focused on the current situation surrounding traceability.

4. *Demonstration*, the artifact is to be demonstrated among the interviewed stakeholders. In this case, this is the distribution of the litepaper in combination with step 5.

5. *Evaluation*, to measure whether the artifact is a good solution, the objectives are compared with the artifact in the context. In our research, we will evaluate the artifact by sending the lite paper to the interviewees accompanied by targeted questions.

6. *Communication*, the last step involves communication of the designed artifact to all the stakeholders, as well as publishing any other related documents, e.g. the research article. This research paper will be published along with the litepaper.

2.2 Literature study

For the knowledge base, we executed a literature study which is part of the DSR framework. This literature study will support the problem identification and setting it design objectives, combined with the exploratory interviews. Due to the broad nature of traceability and automated decision making systems, we will focus on answering the following questions:

- To what degree does the definition of traceability of automated systems apply in the field of Explainable AI?
- What are the existing solutions to ensure traceability of automated outcomes?
- What is the legislation of traceability of automated outcomes?

To find relevant work, we used various search terms focused on the technical side of traceability as well as more policy-oriented work and legislation. Terms we searched with, mostly combined, include but are not limited to:

reproducibility	traceability	lineage
artificial intelligence	explainable AI	insurance industry
auditability	legislation	AI Act

2.3 Interviews- Thematic Analysis

For our research, we will conduct exploratory interviews and a evaluation. To structure our data, we follow a thematic analysis approach. This instantiation is part of the DSR framework, where it covers the environment (social context) and supports defining the problem identification and objectives for a solution.

Exploratory Interviews
To get a broader understanding of the current situation in the Dutch Insurance Industry regarding the concept of traceability, we will conduct an exploratory Interviews. In the exploratory interview, our goal is to elicitate a conversation about topics surrounding traceability such as regulations, practices and deficiencies. The focus lies to researching which traceability requirements apply specifically for the Dutch Insurance Industry (see more at section 5).

Thematic Analysis
For the qualitative analysis of data, we followed a thematic analysis approach. Thematic analysis is a widely used method of qualitative data analysis. Note the following points:

1. Thematic analysis is a flexible method of analysis for seeking to understand experiences, thoughts or behaviours across a data set, e.g. the transcription of interviews. Each transcription is reviewed, after which parts are coded. Codes are specific highlights of a part of the transcription which are based on specific topics, thoughts or findings. A code can be e.g. "uncertainty about the rules", which can re-occur multiple times. With this approach, themes, which are actively constructed patterns, can be derived from the coded parts. In turn, each theme can consist of a feeling or meaning. Finally, these themes can be used for establishing further needs or, following from coded parts of text in the data.

2. For our interview, we will derive these themes inductively. This means that we will create themes and codes following from the data investigated.

3. The data set used are audio transcriptions of the four interviews.

We will use the steps described by Kiger and Varpio (2020) which are as follows:

1. *Familiarizing with the Data.* In this step it is important to know what the data is about. Kiger mentions that with auditive data, transcribing is sufficient to familiarize yourself with the data.

2. *Generating Initial Codes.* This step consists of creating a coding framework in order to code all the data. For this research, coding was done manually, but with assistance of Atlas TI Web.

3. *Searching for themes.* In inductive analysis, themes are derived expressly from coded data. So themes identified will be closely linked to the original data. Here, themes are searched with the help of grouping the codings in different kind of categories. Themes can be mapped with each other in a visual way.

4. *Reviewing themes.* This step consists of re-reading and revising codes and themes. Coded data within each category should have a proper fit and adequately represent the entire body of data.

5. *Defining and naming themes.* Here a narrative description and definition of each theme is created.

6. *Producing the Report.* In this step, a report is produced consisting of the writing process, code and theme selection. It should consists of not only how the researcher interprets the data, but also why certain selection for themes of the data are important and accurate.

3 Related Work

In this section, we will shortly discuss earlier work and research related to the topic of traceability. Hereafter, different parts of each research will be used in our literature study. Note that in this section, we briefly discuss the research one by one. In the section Literature Study (4) we will use the related work and organize it into more topic oriented structure (definition, legislation, requirements, tools, etc.).

Firstly, Kroll (2021) researched and outlined traceability as a principle for operationalizing accountability in computing systems. He examined various ways "how traceability has been articulated in AI principles and other world wide policy documents". With this, he distilled a set of requirements driven by the principle of traceability and systematized various technologies to meet these requirements. Eventually Kroll also identifies "gaps and needs separating what traceability requires and the tools already available".

Another more common work is the document with "Ethics Guidelines for Trustworthy AI" (European Commission, 2019). This document is written by the High-Level Expert Group on AI (HLEG AI) and is focused to support Trustworthy AI by proposing several guidelines, divided into 7 key requirements. It also includes an assessment list with points to assess to obtain a certain level of AI. The HLEG AI argues that Trustworthy AI consists of three components: lawful, ethical and robust. The HLEG AI hopes to set a framework, aiming to offer guidance for ethical and robust AI systems. For the realisation of Trustworthy AI, the HLEG AI identified 7 key requirements where one of which includes transparency. In turn they argue that traceability is a **component** of transparency, enabling identification of reasons why or why not an AI decision was erroneous (see more at 4.2.1). In our research, we will use this work to structure topics surrounding AI and use the assessment list for traceability as umbrella entity.

Based on the document "Ethics Guidelines for Trustworthy AI", the European Commission, Directorate-General for Communications Networks, Content and Technology (2021) drafted the so called "AI Act". This draft legislation elaborates on concretizing harmonised rules on Artificial Intelligence. The European Commission also proposes a regulatory framework on Artificial Intelligence with specific objectives, including its trustworthiness, traceability and record-keeping. In this research, we will frequently refer to this proposal. However, note that the act is still in draft version and details may be changed in the future.

Mora-Cantallops et al. (2021) elaborated further on the notion of the HLEG AI that "traceability is considered as a key requirement for trustworthy AI" and reviewed relevant tools, practices and data models for traceability. They also "propose some minimal requirements to consider a model traceable according to the HLEG AI". The research of Mora-Cantallops et al. (2021) proved useful for identifying various elements for traceability as well as for reproduciblity, eventually placing their findings in practical solution.

Raji et al. (2020) propose an end-to-end framework for internal algorithmic auditing to support AI system development end-to end. They use stages for auditing, where each stage contains a set of documents that can be reviewed or created, eventually leading to an overall audit report. This research, following Raji et al. (2020), intends to ”contribute to closing the accountability gap in the development and deployment of large-scale AI systems”. Note that this research is focused on closing the accountability gap and only partly covers the principle for traceability. Nonetheless shows this research a good overview of what to consider when developing an AI system. We will use this research to identify artifacts which should be documented to proper fit the early stated traceability requirements.

For the more industry specific related work, we will use various works from the AFM & DNB. Firstly the AFM (2018) published a report on its view on the so called ”Robo Advice”. In other words, the AFM elaborates on various points of interest when it comes to advice which originates from automatic decision making. It focuses on legislation, e.g. the Financial Supervision Act (Wet Financieel Toezicht), as well as other important topic such as: audience, explanations, validity period of an advice, customer data and privacy legislation.

Another work of the AFM and DNB (2019) elaborates on the recent developments of AI in the Dutch Insurance Industry. It identifies risks, points of interest as well as opportunities related to the deployment of AI. More specifically, they highlight three technical components namely the input data of a model, the technique used behind a model and the model outcomes. Relating traceability, this work bridges the gap between the (technical) knowledge context and the (industry specific) social context by presenting figures about the deployment of AI and providing points of interest matching the current situation. In our research, we will use this exploration in order to map the questions needed to be asked in order to attain a certain level of traceability.

4 Literature study

To support our knowledge base, we have executed a literature study. Here we firstly expand on the scope of automated decision making systems. Thereafter, its definition, urgency and relations with other concepts. Finally presenting an overview of several focal points to consider to ensure traceability.

4.1 The scope of automated decision-making (ADM)

In this section, we will classify various automated decision-making (ADM) systems, their application and their place in the value chain for the financial industry. We follow the approach of Kroll (2021) and try to avoid the term ”Artificial Intelligence”, which eludes a rigorous definition. Instead of using the term ”Artificial Intelligence”, using the term ”automation” embodies a broader range of technological artifacts. The Association of Insurers (Verbond van Verzekraars) published an Ethical Framework (Verbond van Verzekeraars, 2015) and also avoids the term ”Artificial Intelligence”. The Association maintains the definition ”Data Driven Decision Making” and does not give a specific definition for AI. This is because the Association also follows the reasoning that the scope should be broader than only AI techniques. ”Decision Making” is not only focused to the decision to accept customers or not. It is focused on all decisions that a insurer makes. However, in this section and the next, note that ”Artificial Intelligence” or ”Machine Learning” terminology still will be encountered to remain correct when we are referring to the investigated articles and literature. Even though e.g. the terminology AI is used in the literature, our research focuses on a broader range: ”Automated Decision Making Systems” (ADM System’s).

First, we will classify various ADM systems based on the assessment list for traceability of the AI High Level Expert Group (European Commission, 2019): rule-based systems and learning-based systems. Secondly we will expand these classes into more specific examples of ADM systems structured by AI models based on the view on taxonomy of AI given by Koster et al. (2021), in turn based on the taxonomy of Barredo et al. (2020). Thirdly, we will consider briefly the current developments around the terminology of AI by looking at the definition of AI in the upcoming AI Act (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021). This all together create the scope of ADM systems in our thesis:

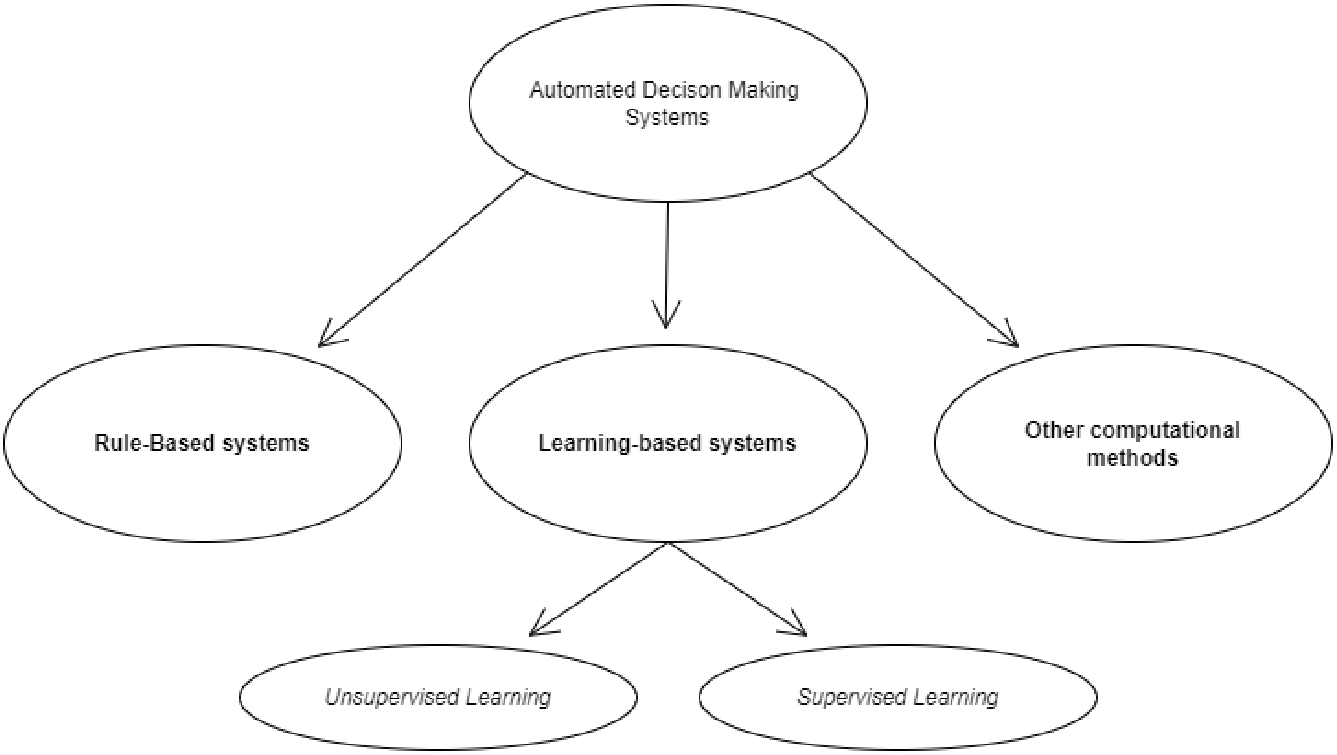


Figure 3: ADM's and the categories used in this thesis

4.1.1 Rule-based and Learning-based AI systems

The AI High Level Expert Group (European Commission, 2019) makes a distinction between learning-based and rule-based AI systems in its assessment list to ensure traceability. As the name already states, rule-based systems are based on a pre-existing set of rules. Learning-based systems are systems that have the ability to learn and adapt new behaviour whilst running. These learning-based systems can be divided into supervised learning methods and unsupervised learning methods. Supervised learning methods are machine learning (ML) methods where the available data consists of labelled examples. Unlike unsupervised learning, where the available data consists of unlabelled examples. We will not go into further technical detail which specific requirements apply for each type of system, however we will use this distinction as well as the corresponding assessment list of the AI HLEG to structure traceability concept's (see more at section 4.3).

4.1.2 Taxonomy of AI models by Koster et al. (2021)

For AI models, Koster et al. (2021) explained for transparent models as well as non-transparent models how they fare with respect to transparency and explainability according to Barredo et al. (2020). Note that explainability is another concept than traceability (see also 4.2.3). To remain consistent, we will use this taxonomy as explained by Koster et al. (2021) to give an overview of which systems we are talking about. Koster et al. (2021) make a distinction between transparent and non-transparent models. Following his argumentation, transparent in this case entails that a model has *simulatability* (the degree of which a model is able to be simulated by a human), *decomposability* (the degree of which a model can be decomposed into individual components) and *algorithmic transparency* (the degree of confidence of a learning algorithm to behave 'sensibly').

Transparent models, described by Barredo et al. (2020) and Koster et al. (2021)

- *Rule-based*, this model is often referred to "business rules" and often follows the concept of "IF x happens, then Y". Rules are decomposed and are algorithmic transparent, however this becomes more and more difficult when the number of rules increases.
- *Linear / Logistic regression*, this model takes the assumption of linear dependence between predictors and the predicted variables. To maintain decomposability and simulatability, its size must be limited and the variables used must be understandable by their users. Variables and interactions are too complex to be analyzed without mathematical tools. Logistic regression is a **supervised learning** classification model.
- *Decision Trees learners*, decision trees are hierarchical structures for decision making, used to support regression and classification models. It uses a decision tree as a predictive model to go from observations (branches) to conclusions (leaves). Decision Trees is a **supervised learning** model.
- *K-Nearest Neighbors*, this model deals with classification problems where it predicts a class of a test sample based on its K nearest neighbors (where the neighborhood relation is induces by a measure of distance between samples). K-nearest neighbours is a **supervised learning** model.
- *General additive models (GAM)* are mainly used to understand relationships between variables in the dataset, rather than predict outcomes with a certain accuracy. The variable to be predicted is given by the aggregation of a number of unknown smooth functions defined for the predictor variables.
- *Bayesian Models* usually takes the form of a probabilistic directed acyclic graph whose links represent conditional dependencies between a set of variables. The model convey a clear representation of the relationships between features and the target.

Non-Transparent models (Barredo et al., 2020)

- *Tree ensembles*, often known as random forest, combine several decision trees to produce better predictive performance. This combination of decision trees loses its transparency. This type of model is a **supervised learning** technique.
- *Support Vector Machines (SVM)* is associated with algorithms which analyse data for classification and regression analysis.
- *Neural networks (NN)*, is a **supervised learning** method which resembles a neural network of living organisms. NN comes in many different shapes. Koster et al. (2021) explicitly discusses three types: multi-layer neural networks also known as multi-layer perceptrons (MLP), convolutional neural networks (CNN) and recurrent neural networks (RNN).

4.1.3 On the definition of AI in the Artificial Intelligence Act proposal

A proposal for an Artificial Intelligence Act is dispatched to the European Council and European Parliament, based partly on the report of the AI High Level Expert Group. Annex I of this proposal for the Artificial Intelligence act (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021) lists techniques and approaches which can be used to develop an Artificial Intelligence system according to Article 3, point 1 of the AI act. This includes:

> ”
>
> 1. Machine learning approaches, including *supervised, unsupervised and reinforcement learning*, using a wide variety of methods including deep learning;
> 2. *Logic- and knowledge-based approaches*, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
> 3. Statistical approaches, Bayesian estimation, search and optimization methods;
>
> ”

Since the proposal is currently under discussion by both the European Council and European Parliament at the time of writing, we will maintain the categories **rule based systems** and **learning-based** systems. Note that point *”1. Machine learning approaches (...) including deep learning”* covers **learning based systems** and point *”2. Logic- and knowledge-based (...) and expert systems;”* covers **learning based** as well as **rule based** systems.

4.1.4 Current use of ADM systems in the Insurance Industry

The AFM published an exploration of Artificial Intelligence in the insurance industry. The AFM and DNB (2019) state that Machine Learning (ML) techniques are, at the time of writing (July 2019), not used structurally or on a big scale. Insurers mostly focuses on expanding existing statistical models and deploy ML techniques more ad-hoc or as support or control on regular models. One specific example of the application of a Machine Learning method is the use of Natural Language Processing: interpreting and understanding text data. However, this is mostly used for back office tasks such as sorting and allocating mails and not mainly used for main processes for insurers. In the value chain, AI applications and ADM systems could be deployed on a broad range of insurance processes such as product development, risk selection, pricing, acceptation, claim management and fraud detection.

We have identified Automated Decision Making (ADM) systems with the use of the assessment list for traceability of the HLEG AI (European Commission, 2019) by using the categories Rule-Based systems (RBS) and Learning-Based systems (LBS). There after, we expanded these categories by giving examples with the taxonomy of AI models as defined by Koster et al. (2021) and Barredo et al. (2020), who also make a distinction between transparent and non-transparent models. Finally, we elaborated on the definition of AI with the use of the AI proposal (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021) and matched these with the taxonomy as well as the two categories RBS and LBS. All this together forms the scope of the ADM systems we are talking about in this research.

4.2 Traceability and its urgency

Now we have established the scope of the systems where traceability applies to, we will discuss the definition and urgency of traceability. With its definition, we will attempt to give a broad overview of why traceability is an important concept and how this relates to guidelines and legislation, eventually leading to its requirements.

4.2.1 Definition

The terminology of traceability is a definition which can be used in various ways. For this research, we will first discuss different angles to define traceability. Next we will operationalize the definition which will be used in this this study. Finally, to exclude confusion, we will also elaborate on the differences and connections between traceability in regard to transparency, explainability, accountability and reproducibility.

Traceability as in ISO 8402
Olsen and Borit (2013), discusses various definitions of traceability, ranging from the general term as well as the definition regarding the food industry. The objective of Olsen and Borit research was to ”examine the use of the term traceability in scientific articles in various contexts”. One of its definitions of traceability Olsen gives is the one of ISO 8402 (“NEN-ISO 8402:1994 nl”, 1994):

> ”The ability to trace the history, application or location of an entity by means of recorded identification”

This definition was withdrawn in 2000, but to this day, it is still commonly used various scientific articles. (Olsen & Borit, 2013)

Traceability in the EU General Food Law
Another definition Olsen and Borit (2013) elaborates on, is a definition which applies to the food industry and is specified by the EU General Food Law:

> ”The ability to trace and follow a food, feed, food-producing animal or substance intended to be, or expected to be incorporated into a food or feed, through all stages of production, processing and distribution.”

This definition literally states ”the ability to trace and follow”, which means that traceability intends to be able to trace an entity (in this case i.e. food) through all stages of production, processing and distribution. Note that this definition focuses on tracing and following the position of an entity and differs from the definition of ISO 8420. Olsen and Borit (2013) argues that ”the definition by the EU General Food Law is less detailed when it comes to describing what types of properties are relevant or how traceability might be implemented”. ISO 8420 however expands on this notion by tracing the history, application or location of an entity by means of recorded identification.

Traceability in global policy documents

The European Commission (2019) set up an independent High Level Expert group on Artificial Intelligence (HLEG AI). The members of HLEG AI presented seven key requirements to attain "Trustworthy AI" in the document "Ethics Guidelines for Trustworthy AI". One of its key requirements is transparency, in turn consisting of three components: *traceability*, *explainability* and *communication*. In its glossary the following definition of traceability is mentioned:

> "Traceability of an AI system refers to the capability to keep track of the system's data, development and deployment processes, typically by means of documented recorded identification."

Note that this definition looks quite similar to the definition in ISO 8402, especially with the phrasing "by means of recorded identification".

The European Commission (2019) also states that traceability is a component of transparency, which is in turn linked as a key requirement to attain trustworthy AI. Here, the following is mentioned about traceability with emphasis added:

> "The **data sets** and the **processes** that yield the AI **system's decision**, including those of **data gathering** and **data labeling** as well as the **algorithms used**, should be **documented to the best possible standard** to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn,could help prevent future mistakes. Traceability facilitates auditability as well as explainability"

Kroll (2021) researched a variety of global policy documents including the EU High Level Expert Group's "Ethics Guidelines for Trustworthy AI", but also policy documents outside the European Union coming from North-America and China. The findings of Kroll (2021) on traceability in these various policy documents typically follows the same reasoning as seen in the AI High Level Expert group's mention about traceability:

> *"Across a variety of global policy documents, then, we see that traceability has emerged as a key requirement for the responsible use of software systems. This property entails systems where the design methodology, underlying data sources, and problem definitions are clearly documented and released to stakeholders"* (Kroll, 2021)

Traceability and its definition in this thesis

With the definitions as explained above, we set boundaries for the scope of this research with the help of the mentions and definitions of traceability in the report "Ethics Guidelines for Trustworthy AI" of AI High Level Expert Group (European Commission, 2019), and the views of Kroll (2021) and Olsen and Borit (2013).

First, we need to underline the word *"data"*. This word is used both in the definition as well as in the mention of traceability to facilitate explainability. It says that traceability refers to the capability to keep track of a system's data. In other words, the capability to keep track of the data sets and processes.

Second, we need to underline *"development and deployment processes"*. In this case, the deployment and development of an AI system. This part of the definition also comes back in the mention of traceability as a requirement for explainability: "AI system's decisions, including those of data gathering and data labeling as well as the algorithms used".

Finally, *"documentation"* will be another important keyword. The European Commission (2019) state that thus both concepts around "data" and "AI development and deployment processes" should be documented to the best possible standard, which allows traceability and increase transparency. This also applies to "AI system's decisions", which enables identification of the reasons why an AI-decision was erroneous (or not).

4.2.2 Urgency

Now we have established a definition for traceability, other important questions arises. To what degree does the definition of tracability extend? Why is this an urgent topic? In this short section, we will consider national (Wet Financieel Toezicht) and European laws (AI act proposal) and regulation which highlight the urgency of traceability in a juridical sense. After which, we will go into the more technical urgency of traceability (AFM & DNB, 2019).

Financial Supervision Act (Wft)

One of the main national laws in the Netherlands is the Financial Supervision Act (Wet Financieel Toezicht / Wft) which is monitored by the AFM. In this Financial Supervision Act, stakeholders are obligated to guarantee compliance with the duty of care (article 4:24a Wft[4]) which entails that the stakeholder should act in interest of the customer. This also entails whenever a contract has been or can be established by e.g. a consumer and insurer, the insurer is obligated to notify the customer of all information needed in order to enable the customer to make an adequate assessment of the financial product or service. This is often called "the duty to inform " and can be found in article 4:19 and article 4:20 of the Financial Supervision Act. Traceability supports this duty inform in a sense that it facilitates the ability to inform whenever requested.

In a vision of the AFM at automated decision making (AFM, 2018), or how the AFM calls it "Robo Advice", no distinction is made between "Robo Advice" and physical advice when it comes to the duty of care. However, special attention should be made when it comes to "Robo Advice". The duty to inform entails that all information in the advice should be correct, clear and not misleading in accordance with article 4:19 Wft. A consumer cannot, unlike physical advice, ask verbal questions about clarification or justification of the advice. Whenever an algorithm is used and it turns out an error occurs, the insurer should have a process to suspend the advice. The advisor should also determine the impact and extent of the identified error and inform and compensate its clients. Traceability enables the advisor to establish the impact, identify where the error occurred and why. Hence careful development of algorithms can be facilitated.

[4] https://wetten.overheid.nl/BWBR0020368/2020-12-29

Whenever a third party provides a product (ADM system) which generate an automated advice, the insurance company who eventually gives out the advice is the one who is obligated to weight all risks and take adequate measures to comply to the duty of care (page 22, 4.1.3. from AFM, 2018). The supplier of the product should have installed a process aimed to monitor the quality of the product. But most the insurance company is accountable for the quality of the given advice and the management of the underlying system. To be able to exercise accountability, the insurer needs to understand the rationale, risks and decision rules of the algorithm. Traceability facilitates auditability and accountability (see 4.2.3). Also in the case of the use of a third party ADM system, traceability supports the ability to monitor the given advices. It also exercises accountability in a sense that every decision or outcome can be traced back to how that certain decision or outcome has been established.

The Artificial Intelligence Act

The Artificial Intelligence Act (European Commission, 2022) is a proposed regulation which aims to introduce a regulatory and legal framework for AI, encompassed for all sectors and all types of Artificial Intelligence. This act protects customers, among other things, e.g. against discrimination by Automated Decision Making Systems. The proposed act aims to provide developers, operators and users with clear requirements and obligations regarding uses of AI. This is done by a risk based approach, where AI systems are divided into unaccepatble, high, limited, minimal or no risk segments. A high-risk AI system will be subject to stricter obligations than a limited-risk AI system.

At the moment of November 2022, a final tweak has been made which is worth mentioning:

> "In a final tweak to the text, only algorithms used for the risk assessment and pricing of health and life insurance are considered high risk. In contrast, the rest remains covered by sectorial legislation. Algorithms used to evaluate individuals credit scores, credit worthiness or insurance premiums were put in in the high risk basket. However, micro or small companies putting into service these systems for their own use have been exempted". (Bertuzzi, 2022)

Regarding traceability, the AI Act mentions documentation and traceability as requirement for high-risk AI systems:

> "In case infringements of fundamental rights happen, effective redress for affected person need to be made possible by ensuring transparency and traceability of AI systems coupled with strong ex post controls" (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021)

Furthermore, *Article 12 Record-Keeping*[5] goes into detail of the capability of enabling automatic recording of events. It says that "the logging capabilities shall ensure a level of traceability of the AI system's functions throughout its life cycle that is appropriate to the intended purpose of the system". Here after, the article provides a few minimum requirements of the logging capabilities such as: period of use, input data, reference database and identification of natural persons.

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Considering this act, the urgency of traceability in the Insurance Market comes forward in two ways. Firstly, in most cases insurance related processes using AI systems or automated decision making are considered as a high-risk AI system by the AI act (especially assessment and pricing of health and life insurances) (Bertuzzi, 2022). This means that stricter rules and requirements apply and ensuring traceability is considered as a requirement for effective redress in the case of infringements. At last, *Article 12 Record-Keeping* especially applies for high-risk AI systems, hence these stipulated clauses also must be met. At the moment of drafting this thesis (November, 2022), the European Commission is now waiting for co-legislators to finalize their positions. After this, inter-institutional negotiations between the European Parliament, European Council and European Commission will be made.

4.2.3 Traceability and its relation to other concepts

Now we have established the definition of traceability and its urgency emerging from legislation and risks, we will elaborate how traceability relates to other concepts. Traceability lies close to other concepts such as reproducibility, auditability, accountability, explainability, replicability and repeatability. In this section we will attempt to rule out confusions and make clear when we talk about traceability or other concepts.

Kroll (2021) examined ”various ways in which the principle of traceability has been articulated in AI principles and other policy documents from around the world”. Kroll summarized that ”traceability has emerged as a key requirement for responsible use of software systems. This entails design methodology, underlying data sources, and problem definitions being clearly documented and released to stakeholders.” Note that this summary is in line with the definition of traceability used in this paper.

Kroll (2021) further argues that traceability encompasses auditability as well as testability of systems both during development and operation. Auditability is the ability to audit: ”to conduct an official examination of the accounts of a business and produce a report” (Cambridge Dictionary, 2022). Furthermore, Kroll (2021) conclude that traceability seeks to relate disclosed information to whom to hold responsible in cases of both failure and success. Here, traceability provides a link between transparency, provenance and accountability.

Kroll (2021) also states that ”traceability is an expansive concept, serving many values both concrete and abstract”. Here ”traceability ties the reasons a system works as it does to its current operation”. Again, this supports audits and interrogation, in turn serving and operationalizing accountability. But traceability also provides a path to understand systems integrity both in systems of complexity and security.

> ”If an adversary were to substitute or modify components of or inputs to the system for some or all decisions, robust traceability would require that this manipulation become visible to affected parties” (Kroll, 2021)

Related, a traceable system encompasses aspects of explainability: ”a traceable system must be understandable to humans who are intended to trace its operation” (Kroll, 2021). Finally, traceability serves to make plain the reasons behind failures. This shows where and which design choices investigators can make an analysis of once an undesired behaviors occurs.

Traceability and Reproduciblity

Traceability is commonly confused with other terminology such as repeatability, replicability and reproducibility. Mora-Cantallops et al. (2021) state that traceability intersects with these concepts. They mention the definitions of repeatability, replicability and reproducibility stated by the Association of Computing Machinery (ACM) as follows:

- Repeatability (Same team, same experimental setup), a researcher can reliably repeat his/her own computation.
- Replicability (Different team, same experimental setup), an independent group can obtain the same results using the authors own artifact's.
- Reproducibility (Different team, different experimental setup), an independent group can obtain the same results using artifacts that they develop completely independently.

Gundersen and Kjensmo (2017) also argue for a distinction of replication and reproducibility. ”Replication is seen as re-running the experiment with code and data provided by the author”. Note that this is in line with the definition given by the ACM. We will use the definition, which is in line with Gundersen and Kjensmo's definition, for reproducibility in this research:

- *Reproduciblity* in empirical AI research is the ability of an independent research team to reproduce the same results using the same AI method based on the documentation made by the original research team. (Gundersen & Kjensmo, 2017)

Gundersen and Kjensmo (2017) also proposed three different degrees of reproducibility, where an increased degree suggest in an increased generality of an AI method:

- ”R1: Experiment Reproducible. The execution of the same implementation of an AI method produces the same results when executed on the same data. Everything required to run the experiment is needed to reproduce the results.”
- ”R2: Data Reproducible. An experiment is Data Reproducible if an alternative implementation of the AI method is used, executed on the same data. It requires only the method description and exactly the data. However, differences in software and hardware could have significant impact on software.”
- ”R3: Method Reproducible. Here the execution of an alternative implementation of the AI method produces the same results executed on different data.”

In later research from the same author Gundersen et al. (2018), it is argued that R2: Data Reproducibility is often called replicability. In turn Mora-Cantallops et al. (2021) argues that the guidelines for trustworthy AI by the HLEG AI is concerned about assuring replicability. For this research, we will consider replicability (R2 Data Reproducible) as an important part of traceability. This will be elucidated in ”Requirements and focal points to ensure traceability” [4.3].

4.3 Requirements and focal points to ensure traceability

In the previous section, we elaborated on the definition of traceability, its urgency and its connection with other concepts. In this section we will discuss various requirements and practical focal points needed in order to obtain a certain level of traceability. Because of the upcoming AI act, based on the Ethics Guidelines for Trustworthy AI by the HLEG AI, we will use the assessment list of the HLEG AI as umbrella topic in order to structure various steps to execute traceability technically.

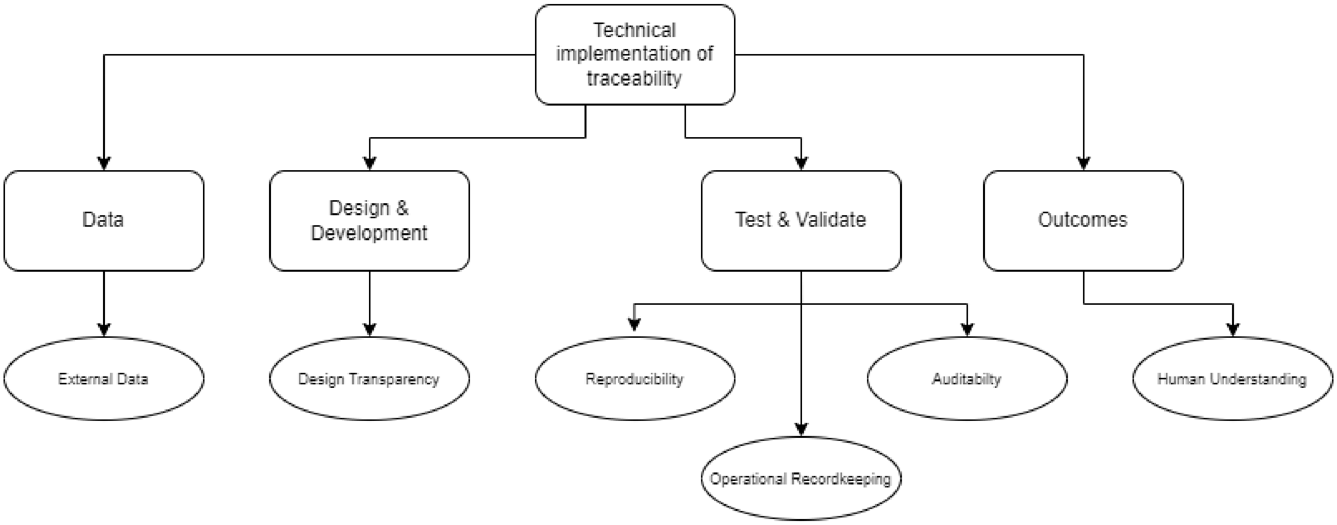


Figure 4: Topics to consider when ensuring traceability, based on the HLEG AI's assessment list.

The AI High Level Expert Group of the European Commission (2019) published a non-exhaustive assessment list in the report ”Ethics Guidelines for Trustworthy AI” to operationalize the 7 requirements for trustworthy AI. This assessment list is primarily addressed to developers and deployers of AI systems and compliance with this assessment list is not evidence of legal compliance. However, it gives a good overview of various components needed to be documented. For the requirement of transparency with its component *traceability*, the questions to assess are the following:

> ”
>
> Did you establish measures that can ensure traceability? This could entail **documenting** the following methods:
>
> Methods used for designing and developing the algorithmic system:
>
> - Rule based AI systems: the method of programming or how the model was built;
> - Learning based AI systems: the method of training the algorithm, including which input data was gathered and selected, and how this occured;
>
> Methods used to test and validate the algorithmic system:
>
> - Rule-based AI systems; the scenarios or cases used in order to test and validate;

- Learning based model: information about the data used to test and validate

Outcomes of the algorithmic system:

- The outcomes of or decision taken by the algorithm, as wel as potential other decisions that would result from different cases (i.e. for other subgroups of users).

”

Note that this assessment list does not specify how to document these methods and outcomes, but only what in a very abstract way. For example: *”Learning based AI systems: the method of training the algorithm, including which input data was gathered and selected, and how this occured”* does not specify any minimum requirements or explanations to document, in this case, i.e. input data. Taking this into account, this assessment list still does give a good generic overview of what to document: ”methods for designing, methods to test and validate and the outcomes of algorithmic systems”.

Kroll (2021) expanded on the notion of traceability he reviewed in various policy documents, including the published report ”Ethics Guidelines for Trustworthy AI” from the HLEG AI. Based on these various policy documents, Kroll gives a broader set of requirements driven by the conceptualization of traceability in these documents. Kroll clearly states that ”these requirements should be viewed as a lower bound. Other activities depending on the application context may be required”. He divides this into 5 different areas: Design Transparency, Reproduciblity, Operational Recordkeeping, Human Understanding and Auditabiltiy. Besides the requirements stated by Kroll, several proposals exists about requirements and artifacts to be documented in order to adhere to the principle of traceability and accountability. Raji et al. (2020) for example, ”introduced a framework for algorithmic auditing that support artificial intelligence systems end to end”. The authors suggest artifacts, such as a set of documents, that form tall together an audit report. Raji et al. claims that ”this proposed auditing framework is intended to contribute to the accountability gap in the development and deployment of large scale artificial intelligence systems”. We will also try to follow this approach by suggesting artifacts as well as topics to consider by implementing traceability. Firstly, we’ll go over the topics and artifacts to consider in the design and development stage. There after, Test and Validation is considered and finally, outcomes as a whole is considered.

4.3.1 Design and Development

The first category which should be considered, are measures which can be established for documenting methods used for designing and developing an algorithmic system. Transparent and traceable documentation in the design phase about how and why an ADM is built is necessary to call out intended and unintended misuse (“ABOUT ML Reference Document, Research Themes on Documentation for Transparency”, 2021). Kroll (2021) argues that for design and development: ”design choices should be made available to stakeholders affected by the system’s operation, which could be accomplished via transparency, such as making system source documentation, code or data available”. However, ”just” making source code and its documentation available” is not an obvious option. Most of the time stakeholders, such as consumers or judges, do not have the knowledge to simply interpret available code or business rules. Not only the amount of code or business rules used will an issue, but also the copyright and privacy constraints. To tackle this issue, we will consider other approaches for documentation, rather than simply "making development and design sources available".

Requirements and disclosures

Kroll (2021) argues that design proceeds from requirements because traceability asks which and why certain design choices were taken. Requirements, often specified in a *requirement document*, describe the systems function and how the system's goals are implemented by the designers. Whenever requirements are not specified formally Kroll argues that there should at least be a "written and ambiguous source in natural language, where the development proceeded from". Many models or system that are a result from exploratory or trial and error processes (e.g. data science practices), must be added with semantic descriptions that accompanies that process. This includes the reasons why a particular choice was or was not made, as well as how certain choices are deemed worthy.

In the development process, collecting documentation mainly implies collecting the documentation of: the record of data, model dynamics, design documents and other product development artifacts. However at times documentation, especially privacy sensitive artifacts, cannot be made available easily. Documentation can also be distributed across different teams and stakeholders (Raji et al., 2020). Also in the case of rule based systems due to the huge amount of rules, documentation can lose over sight and become not understandable.

In this certain cases, retroactive documentation can be enforced with disclosed artifacts. E.g. Raji and Yang (2019) proposes some key artifacts.

- *Design Checklist*. This checklist is a method of taking inventory of all expected and generated documentation in the development process. It ensures which scope of the expected system as well as its documentation needs to be considered, enabling traceability in case of an audit.

- *Model cards & Datasheets*: Two recent standards for more auditable documentation are model cards & data sheets. Model cards is a recent standard explored in a Google Research paper which leveraged to create audible documentation. Mitchell et al. (2019) suggests that models can be documented with their detailed performance characteristics in a so called "model card". A model card, as described by Raji et al., "should include information how the model was built, its assumptions and what type of behaviour it experienced with certain groups of data". "A robust model card is key to documenting intended use of a model, as well as information about evaluation of the data, model scope and risk, and what might be affecting model performance" Raji et al. argues. He also mentions that Model cards are intended to complement "Datasheets". We will elaborate on Datasheetes for Datasets in the data section (4.3.2)

Fact sheets

Another disclosed format to document ADM systems we like to elaborate on are Fact Sheets. Fact sheets (Arnold et al., 2018) focuses on increasing trust in AI services by suggesting a comprehensive set of declaration items tailored to AI. Arnold et al. (2018) envisions that Fact Sheets help to increase trust in AI services by documenting purpose, performance, safety, security and provenance of information to be completed by AI service providers for examination by consumers. A Fact Sheet contain sections on all relevant attributes of an AI service. The Fact Sheet is composed into several categories: statement of purpose, basic performance, safety, security and lineage (traceability). The following questions, compiled by Arnold et al. (2018) aim to overview how a provider keeps track of details that might be required in the event of an audit by an third party. Note that the list below functions as an example and does not include all the questions from a Fact Sheet. The list is taken as example one-to-one from Arnold et al. (2018).

> " **Does the service provide an as-is/canned model? Which datasets was the service trained on?**
>
> - *List the training data sets*
> - *Where there any quality assurance processes employed while the data was collected or before use?*
> - *Were the datasets used for training built-for-purpose or were they re-purposed/adapted? Were the datasets created specifically for the purpose of training the models offered by this service?*
>
> **For each dataset: Are the training datasets publicly available?**
>
> - *Please provide a link to the datasets or the source of the datasets*
>
> **For each dataset: Does the dataset have a datasheet or data statement?**
>
> - *If available, attach the datasheet; other-wise, provide answers to questions from the datasheet as appropriate [to insert citation]*
>
> **Did the service require any transformation of the data in addition to those provided in the datasheet? Do you use synthetic data?**
>
> - *When? How was it created? Briefly describe its properties and the creation procedure.*
>
> " (Arnold et al., 2018)

Versioning
Early development methodologies followed a "waterfall" process where a linear path from design to deployment was followed. Kroll (2021) states that due to often diversions along this path or changes in the environment by the time a system was deployed, "iterative" modalities of software development emerged and versioning control became more and more important. For this kind of development, Kroll (2021) argues that version control of digital artifacts like code, data sets and model products can be enhanced with powerful general versioning-focused methodologies. Gundersen et al. (2018) also recommends to use versioning focused methodologies such as source code repositories to enhance reproduciblity. Gundersen et al. recommends code repositories e.g. general repositories such as GitHub, or language specific repositories such as CRAN for R code.

4.3.2 Input Data

The second category which should be considered, is the input data. In other words, how is this input data gathered and selected and how did this occur? Note that in the assessment list of the HLEG AI, input data is explicitly stated at documenting the learning based models. In this section, we will be focusing on practical methods to document and disclose datasets used.

Datasheets for datasets
Data provenance has been rarely discussed in the machine learning community (Gebru et al., 2018). No standardized procedures currently exists for documenting machine learning data sets. To address this gap, Gebru et al. (2018) proposes *datasheets for datasets* which is "inspired from the electronics industry where every component is accompanied with a datasheet describing its operating characteristics, test results, usage and other information" (N.B. the difference between factsheets is the focuses of the documentation: factsheets focuses on the whole ADM system whereas datasheets for datasets focuses on datasets).

Gebru et al. (2018) states that datasheets aims to increase transparency and accountability, mitigate unwanted biases and facilitate greater reproduciblity of machine learning results. By answering a set of questions, information that a datasheet for a dataset might contain is elicited. "It is grouped into sections, which matches roughly the dataset life cycle: motivation, composition, collection, preprocessing/cleaning/labeling, use, distribution and maintenance" (Gebru et al., 2018). Datasheets have begun to pilot in organisations such as Microsoft, Goole and IBM.

4.3.3 Testing and Validation

For testing an validation, the HLEG AI states that methods used to test and validate an algorithmic system should be documented. For rule based sytems, this could include scenarios or cases in order to test or validate the system. For learning-based systems, it includes information about the data used for test and validate.

Testing
Kroll argues that the gap between what is known to the developers and what is known to the stakholders should be minimized. When testing a model or algorithmic system, a system should have and release information about their test and evaluation plans. To minimize this gap, it necessarily requires disclosing information about the design as well as information about the system's performance under test (Kroll, 2021). Also the HLEG AI states that methods used to test and validate a system should be documented. In specific for rule-based systems, the scenarios and use cases in order to test and validate. In most cases this could be a set of documents showing the concrete rules used in a certain case. But when there are many rules applied, men can lose oversight and solely a document containing all the rules can be to big. In this kind of cases, specific scenario's should be given with trace links between each step.

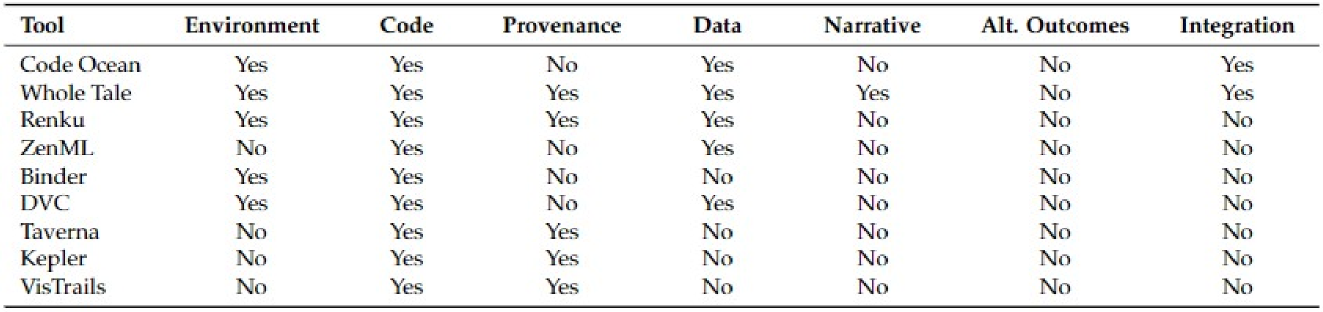
Validation - Reproducibility
To enhance validation, Kroll argues the following about Reproducibility:

> ”It should be possible to reproduce even abstract conclusions from data or any reported experimental results that claim scientific or important authority. If a system’s behavior cannot be reproduced, its intentions cannot be made plain to an outside stakeholder. With reproducibility, an external stakeholder can verify why and how a system was designed a particular way or what a system will be able to do. A risk is the loss or modification of development information (code, data, built components). Thus, robust reproducibility of both artifacts and conclusions must be a requirement for any traceable system.”

A study by Olorisade et al. (2017) identified a set of factors that effect reproduciblity based on reproducing six studies proposing text-mining techniques. The key aspects to facilitate reproducibility are listed below.

- *Dataset*: Information about the location and retrieval process of the dataset is needed. This could be a data repository.
- *Data preprocessing*: The process of ridding the input data of noise and encoding it into a format acceptable for the algoirthm. An independent party should be able to follow and repeat how the data was preprocessed.
- *Dimensionality reduction*: In text mining feature vectors (features translated into numerical vectors) from the preprocessing process, are usally large and sparse. Therefore, the details of the dimensionality reduction technique(s) should be provided alongside the output details to allow for comparison.
- *Dataset Partitions*: Details of how the dataset was divided for use as training and test data (and sometimes validation data).
- *Model Training*: The process of fitting the model to the data. It is crucial to make as much information available as possible. Necessary information include: 1) study parameters, 2) proposed technique details (codes, algorithms etc).
- *Model assessment*: Measuring the performance of the model trained. Here also proposed technique details should be documented to the best possible way.
- *Randomization control*: Most operations of machine learning algorithms or other algorithms involves randomization. Hence it is essential to set seed values to control the randomization process in order to be able to repeat the process.
- *Software environment*: Software packages and modules are in continual development. That is why it is important that details of the software environment used (modules, packages, version numbers, OS) be made available.
- *Hardware environments*: Sometimes, some outcomes can only be reproduced on environments which can handle e.g. large data volumes. Therefore, hardware information is sometimes essential.

Many existing tools exists which aim to support ”methods reproduciblity” or also called ”replicability” of AI systems. Mora-Cantallops et al. (2021) compared a list of tools based on various aspects the tools capture, which is shown in Figure 5. The horizon header displays the various aspects, such as: environment, code provenance etc. In the left side column, the name of the tool is named.

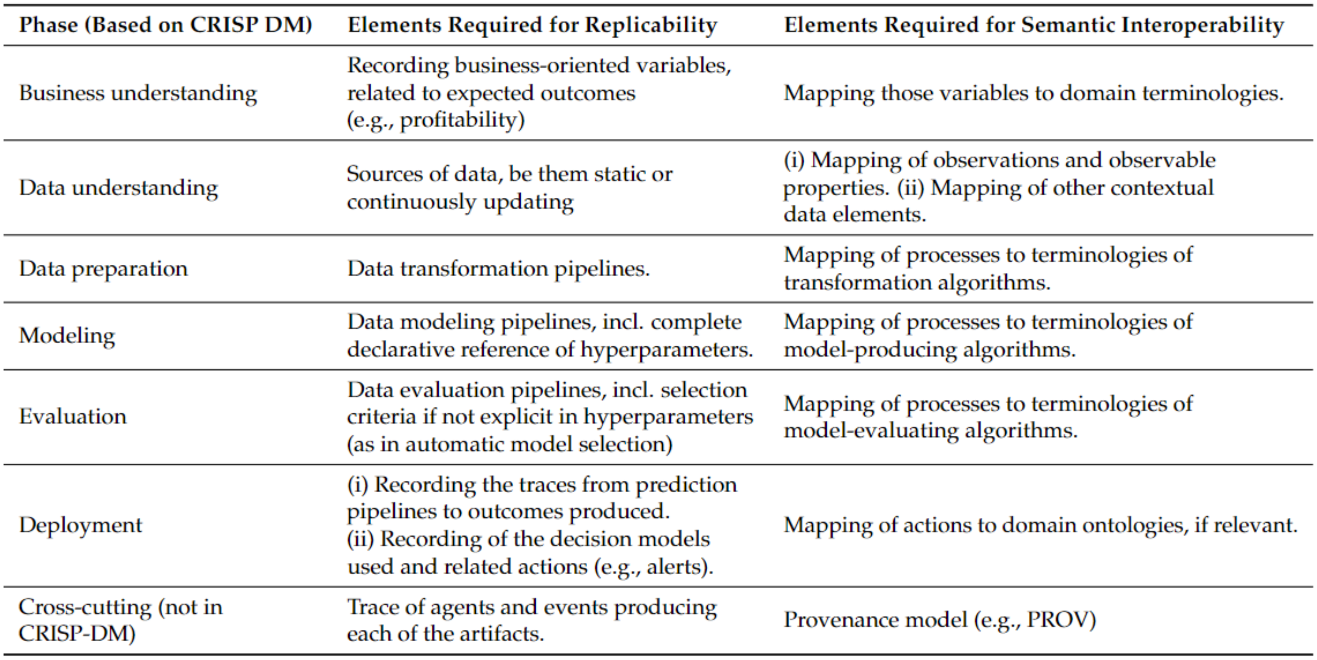


Tool	Environment	Code	Provenance	Data	Narrative	Alt. Outcomes	Integration
Code Ocean	Yes	Yes	No	Yes	No	No	Yes
Whole Tale	Yes	Yes	Yes	Yes	Yes	No	Yes
Renku	Yes	Yes	Yes	Yes	No	No	No
ZenML	No	Yes	No	Yes	No	No	No
Binder	Yes	Yes	No	No	No	No	No
DVC	Yes	Yes	No	Yes	No	No	No
Taverna	No	Yes	Yes	No	No	No	No
Kepler	No	Yes	Yes	No	No	No	No
VisTrails	No	Yes	Yes	No	No	No	No

Figure 5: Comparison between tools that aim to support “methods reproducibility” research (Mora-Cantallops et al., 2021).

Overall Mora-Cantallops et al. state that ”more complete tools cover well the technical side of replicability (environment, code, data, and provenance)”. However, narrative seems to receive less attention. Narrative entails providing detailed information about textual description of the motivation about selecting a particular set of data, model construction or testing. Capturing the environment entails e.g. being able to replicate and capture the OS or software environment. Additionally, no tool brings focus on potential alternative decisions, which is explicitly considered in the HLEG AI guidelines. Finally it is worth noting that some tools are no longer supported or frequently updated. (Mora-Cantallops et al., 2021).

Main elements required for traceability, if aimed for AI systems that are fully replicated, is discussed by Mora-Cantallops et al. (2021). According to the authors, ”the table may serve as a guide for a minimal set of requirements for tools and frameworks”.



Phase (Based on CRISP DM)	Elements Required for Replicability	Elements Required for Semantic Interoperability
Business understanding	Recording business-oriented variables, related to expected outcomes (e.g., profitability)	Mapping those variables to domain terminologies.
Data understanding	Sources of data, be them static or continuously updating	(i) Mapping of observations and observable properties. (ii) Mapping of other contextual data elements.
Data preparation	Data transformation pipelines.	Mapping of processes to terminologies of transformation algorithms.
Modeling	Data modeling pipelines, incl. complete declarative reference of hyperparameters.	Mapping of processes to terminologies of model-producing algorithms.
Evaluation	Data evaluation pipelines, incl. selection criteria if not explicit in hyperparameters (as in automatic model selection)	Mapping of processes to terminologies of model-evaluating algorithms.
Deployment	(i) Recording the traces from prediction pipelines to outcomes produced. (ii) Recording of the decision models used and related actions (e.g., alerts).	Mapping of actions to domain ontologies, if relevant.
Cross-cutting (not in CRISP-DM)	Trace of agents and events producing each of the artifacts.	Provenance model (e.g., PROV)

Figure 6: Guide for a minimal set of requirements for tools and frameworks required for traceability and replicability (Mora-Cantallops et al., 2021).

Validation - Auditability

"Another component of the traceability principles is that they support auditability of systems" (Kroll, 2021) (4.2.3). First, the system must maintain sufficient records during development and operation. This must be done in a way so that "their creation can be reliably established and reproduced" (Kroll, 2021). The task of the term audit is using the traceability principles more in line with assessment. In accounting literature, an audit compares recorded evidence to reality and policies to determine whether that evidence is reliable. Kroll (2021) states the following about assessments and audits:

> "An assessment is the ascription of some value to that evidence or a judgment about the meaning of that evidence. Impact assessments are valuable insofar as they enable traceability, and considering the requirements of traceability can help establish the scope of appropriate impact assessment. " (Kroll, 2021)

Validation - Operational Record keeping and Pipelines

Traceable systems are required to maintain records of their actions and agents. However, which specific records to keep, the duration and the manner of keeping, are all contextual questions. Involvement of industry specific experts as well as subject matter experts who have knowledge of the technology and the industry is necessary to start a valuable record keeping structure.

Transparency could be disclosed through the intervention of trusted oversight entities. However, Kroll (2021) also argues that it is possible to use tools from cryptography to bind the contents of records to the computations performed on those records, when portions of retained information must remain secret or private. An alternative model is to vest record keeping and data access in a trusted third party.

The AI Act (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021) includes an article about "record-keeping" *(Article 12)*. In short, it says that:

> "AI systems shall be designed and developed with capabilities of enabling automatic recording of events. These logging capabilities shall ensure a level of traceability of the AI system's functioning throughout its lifecycle that is appropriate to the intended purpose of the system." (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021)

A minimum is mentioned where high risk AI systems shall provide at minimum:

- Recording of the period of each use of the system
- The reference database against which input data has been checked by the system
- The input data for which the search has led to a match
- The identification of the natural persons involved in the verification of the results

Kroll (2021) argues that traceability requires record keeping both at the development and operational stage. Here the trade off between "direct policy enforcement via the design and compliance with specifications and the detection of policy violations via record keeping and review should be considered". Structured logging is a well studied problem in computing. Kroll (2021) states that many authors have called for the use of structured record keeping tools. Along with traditional software engineering an project management disciplines (life cycles), Kroll says that tools for reproducibility and record keeping comprise a set of currently available techniques which improve traceability but are also currently underused.

4.3.4 Outcomes

The AI HLEG assessment list for traceability by the European Commission (2019) explicitly state: "Documenting the outcomes of or decision taken by the algorithm, as well as potential other decisions that would result from the same cases".
Traceability requires that systems be transparent, not just about their function, but also about whether that function is appropriately communicated to operators and other affected humans. It is important that all the registered documentation is understandable for (inexpert) stakeholders.

Mora-Cantallops et al. (2021) calls this "Semantic Interoperability", or in other words the understandability between the elements of requirements of traceability, the outcomes and the corresponding industry specific context. In the case of Dutch Insurance Industry, technical elements of the development and deployment process, as well as outcomes from ADM systems should be mapped to understandable an relevant terminology for this specific industry.

Because outcomes also relate to the concept of explainability, we will refer to the research of Koster et al. which goes into further detail on explainability of the outcomes of Automated Decision Making Systems.

5 Exploratory Interviews

5.1 Stakeholders of the Dutch Insurance Industry

The Dutch Insurance Industry consists of different parties and actors who do business with each other (De Stichting Vervoeradres, 2022) :

- *Insurance Companies*, mostly "Naamloze Venootschap" or Limited Companies, take out insurance policies for their own account and risk. Besides "normal" insurance companies, there exist proxy companies and agents. These proxy companies underwrite insurance, on behalf and for the account of an insurer.
- *An insurance broker* is a proxy company that mediates the conclusion of insurance policies between policyholder and insurer at (big) insurance exchanges. A characteristic of an insurance broker is that the broker consults with a policyholder on the terms of an insurance policy and submits it to the insurer.
- *An insurance intermediate* is a proxy company that mediates insurance policies between policyholder and insurer in the private market.
- *An policy holder* is a consumer or a company that purchases an insurance from an intermediate or insurance company.

Besides these stakeholders, there are two supervisory bodies and various non-binding organizations who regulate and audit the insurance companies. The two supervisory bodies are the "Autoriteit Financiële Markten - AFM" and the "Nederlandsche Bank - DNB". DNB focuses on solid and trustworthy financial institutions to adhere to obligations. AFM is responsible for behavioural supervision. Both conduct audits and test and work closely together.

Another important actor is *SIVI*, which is an independent research and standardization institute for financial services. SIVI develops and adminster standards for digital business in the insurance industry and more.

To get a broader understanding and familiarity with the insurance domain, we conducted 4 interviews with 3 kinds of stakeholders:

- *A major insurance group (A)*, consisting of around 4000 employees and offers a range of services varying from insurance to mortgages. With this stakeholder, we interviewed a data scientist as well as an ethicist in two separate interviews.
- *A smaller insurance intermediate (B)*, consisting of around 100 employees. Offering mostly a range of business insurances. Here we interviewed a manager of operations and a software developer.
- *A software vendor (C)*, consisting of around 50 employees. Offering mostly software solutions for the financial industry. Here we interviewed a lead business analyst, which was an expert on the application of their products as well as the technical side of development.

The goal of our interviews is to get an understanding of the current state of the knowledge, policies and application of traceability guidelines in the insurance domain. Please note that these interviews focuses on exploring the needs and demands around traceability, where we will eventually be looking for a base for a practical solution. To support this, topics in the interviews are focused on the following:

- What kind of automation methods are used, for what purpose and which goals?
- Do these parties have (extensive) knowledge of the existing guidelines for ”Trustworhty AI” by the European Commission HLEG AI as well as the guidelines from ”Het Verbond van Verzekeraars”?
- How are these guidelines (officially) implemented in the organization?
- Did these parties experience situations where traceability, as requirement, was needed in order to meet a certain demand?
- What kind of gaps, deficiencies or questions do these parties have in order to meet the requirements for traceability?

The first contact with these parties was done by SIVI. Hence forward, interviews were planned with a interview time of 45 minutes. In all cases, this time was sufficient to get enough information. For our research, we followed a thematic analysis approach as defined by Kiger and Varpio (2020). For familiarization with the data, we transcribed the recorded interviews to text. Here we already get a little understanding about the content of the the interviews. After all the interviews were transcribed, we started coding. We derived some code’s such as: ”third party involved”, ”external data”, ”lack of knowledge or expertise”, ”unclear who is responsible”. Eventually we coded the 4 interviews and used the report function to see whether a code came back frequently. An example of a code which came back very often was ”lack of knowledge or expertise”, especially regarding the regulation of traceability. With this, we created a few themes and topics which returned often in all 4 interviews. After this, we reviewed all the data again an re coded some parts when necessary. Eventually we defined themes with a narrative description in a working document. Instead of producing a standalone report, the findings can be found in 5.2.

5.2 Findings

To answer the questions in the interview, we will go through each question and the response of each company. Each person and/or company is labeled with A, B, C (see stakeholders at the beginning of 5.1).

Automation methods used
All companies do use different automation methods. Company A uses the most broad range of automation methods from NLP, Dynamic Pricing to Fraud Detection. Their developed methods are used in their so called business line. Company B started using soft rules for its automated decion making processes and uses this mostly to this day. They also uses Fraud Detection modules from third parties (FRISS). Company C, the software vendor also mostly uses a business rule engine for their software.

Knowledge of traceability guidelines

- Company A's data scientist was not familiar with the guidelines of the AI High Level Expert Group from the European Union. However, its data scientist was familiar with the guidelines and tried to translate/transpose various ethics guidelines to understandable working documents.

- Company B did not really specify whether they where familiar with these guidelines or not. However they specified that they lack people who have enough knowledge to practice traceability.

- Company C was not familiar with the guidelines, but do have a development process.

Implementation of traceability guidelines

- As already mentioned, company A's data scientist was not familiar with the guidelines. They do have best practices to comply to various requirements for explainable AI, including traceability. They are not official policy documents, but best practices which everyone adheres to as best they can. The ethicist of company A said that the topics surrounding explainable AI was very distributed over the company. Right now, they are working on getting to land all those topics on different places in the organization.

- Company B do not have official policy documents. They do think it is a nice development that models are developed for this, but they also consider that it can lead to over-regulating.

- Company C, the software vendor, adheres mostly best practices for developing, testing and implementing machine learning applications. They do have a development process and working documents, but not official guidelines or policies to account for traceability.

Situations where a result had to be reproduced
All companies did not encounter a situation where a result had to reproduced which was resulting from an arising complaint, request or obligation. However, they all claim that they can reproduce an outcome if needed.

- Company A keeps scripts and data in versioning control hubs. They add inline commits about the choices made and explain more about the code / alternation in their commits.

- Company B uses XML data/documents to explain their choices for certain decisions/outcomes which they add to their customer file.

- Company C elaborates on the business rules used and also provides the opportunity for inspection if requested.

Other gaps, feelings, deficiencies or questions regarding traceability
Below, we will list other gaps, deficiencies and questions found which we think is worth mentioning. Here we will not specify each company if it is a theme that recurs often. If a certain theme only appears in one company, we will of course mention this.

- *Accountability internally*, companies are also confident that the results generated by an automated system can be accounted for by another human. How they do this in a practical sense varies from each company.

- *Accountability externally*, a issue that emerged is the one of accounting for external data and outcomes. For example, if you use external data sources or third party applications.

- *Lack in knowledge and expertise.* There is often uncertainty as to whether a company has sufficient knowledge or expertise about the topic of traceability. Specific guidelines, minimum requirements or other policies are not always clear: "when is sufficient, sufficient enough" - Company A.

- *No standardization.* There is lack of standardization and official policy documents. Right now, best practices for the machine learning processes are used. "We think we come very far whilst not using official policy documents or standards".

- *Need for base set of compliance and practical solutions.* There is a call for a base set of compliance and useful grips in order to counter the lack of knowledge and lack of standardization.

6 Results

In this section, we will present the results from the Design Science Research process. First, we will dive into the problem identification and identify why traceability of outcomes is an urgent topic and which problems arise. Second, we discuss the results from the exploratory interviews, and map these into objectives for a solution. There after, we will go into the creation of the artifact and underpin the artifact with the found literature and objectives. Lastly, we discuss the results of the validation process, based on the work of Riemenschneider et al. (2002).

6.1 Problem identification

The identification of the problem starts with the current situation of traceability. This resulted in the presentation of several industry specific cases that have shown the need for sufficient traceability. Not only these cases stipulate the urgency for traceability, but also various authors discuss the urgency and argue that, to tackle such cases, there is a need for practical mechanisms to ensure traceability and move beyond guidelines.

AFM and DNB (2019) state that ”an insurer should be able to identify the relationship between the outcomes, the model and the input parameters” and Brundage et al. argue that standard setting bodies should work with academia and industry to develop traceability (audit trail) requirements for safety critical applications (such as insurancy) of AI systems.

Also regulations and guidelines, such as the Financial Supervision Act and the AI-act, resulted in the identification of the urgency of traceability. In the Financial Supervision Act, its urgency comes forward in compliance where insurers are obligated to comply to the ”duty to inform” as well as the ”duty of care”. Here traceability is needed to support these obligations. It enables an advisor, in case of an error, to establish the impact, identify where and why the error occurred.

The AI act states that especially with the use of high risk AI systems (which is the case in the insurance industry) traceability should be considered as a primary requirement. Here, in case infringements of rights happen, effective redress can for the affected person can be made possible by ensuring traceability. (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021)

In conclusion, the problem identification arise from various sources. Industry specific sources, regulations as well as earlier research identify that there is a need for practical mechanism which elaborate on the requirements for traceability.

6.2 Objectives for a solution

The exploratory interviews resulted into 3 main objectives which can be implemented a potential solution (see more at 5):

1. *Clarification of guidelines and regulations.* Some interviewees were familiar with guidelines for traceability, but most of them were not. That is why the first objective should be the

clarification of the existing rules and guidelines which apply when trying to guarantee traceability.

2. *Base set of requirements.* What are the requirements for traceability? When is good, good enough to comply to these traceability requirements? This objective is to establish a base set of requirements, based on the need for structure and oversight.

3. *Preventing over-regulation.* The interviewees have indicated that there is a call for a practical solution, but fear the issue of over-regulation. The objective we have established is that the artifact should be a practical hands on solution, without restricting stakeholders to their own needs.

Here objective (1) clarification of guidelines and regulations as well as objective (2) base set of requirements can be achieved without violating objective (3) preventing over-regulation.

6.3 The Artifact

Looking for the best fit with the stated objectives, we have chosen a litepaper as artifact. A litepaper is a document designed to compactly set out a topic or information. This means that a litepaper should be easy to read, not too technical, and provide support for the reader's daily work. It is meant for guidance for developers, as well as an elaboration for non-technical stakeholders.

The litepaper is designed to clarify the requirements and guidelines for traceability from various instances. There after, various points of interest in the categories Design & Development, Testing and Validation and Outcomes are discussed together with corresponding questions, explanations and examples.

6.4 Demonstration

The demonstration of the artifact is done by emailing the litepaper to the interviewees together with a response form. In total the litepaper consists of 20 pages and is written in Dutch. We have included the litepaper in the appendix of this research.

6.5 Validation

For the validation of the artifact, we follow an approach based on the work of Riemenschneider et al. (2002) who examined and compared five theoretical technology acceptance models. An acceptance model is a model where various constructs and determinants are proposed to validate the acceptance of software. In the research of Riemenschneider et al., usefulness, voluntariness, compatibility and subjective norm (social factors) were found significant as determinants of the intention to use the reviewed technology or software. Beside these four significant determinants, we will use two more measurement scales (behavioral intention and ease of use) to compile a set of questions for the validation of the artifact. In this section, we will present the questions, as well as the results on the response of these questions. For the responses, the number in the parentheses are the given scale with 5 as the highest (e.g. most likely to disseminate the litepaper, most useful) and 1 the lowest (e.g. less user friendly, less appreciated).

Behavioral intention

- On a scale of 1 to 5, to what extent do you plan to disseminate and use the litepaper in your organization?
    - *(3): Interviewee A states that due to the ongoing implementation of ethics guidelines, the artifact would definitely be taken into account.*
    - *(4): Interviewee B states that they will disseminate the litepaper to selected receipents in the organisation.*

Usefulness

- On a scale of 1 to 5, to what extent does the lite paper contribute to new knowledge about traceability requirements?
    - *(5): Interviewee A, who did not had a lot of technical knowledge about traceability, states that the information in the litepaper is new and useful surely useful.*
    - *(4): One interviewee states that in particular, the litepaper helps to raise awareness about traceability and make it of the development process.*
- On a scale of 1 to 5, would this litepaper make it easier to do your daily work?
    - *(2): Interviewee 1, who is an ethicist, states that he will consult data scientist colleagues for the application and interpretation of the litepaper in daily practice.*
    - *(3): Interviewee 2 states: "We already perform many steps in our process that ensure traceability and acknowledge steps as well. But in terms of documentation regarding traceability, we could still make improvements."*
- On a scale of 1 to 5, how useful do you find this litepaper?
    - *(3): For its usefulness, interviewee 1 can not yet determine to which extent the litepaper will be used, but any best practices and examples in this area may prove useful at some point.*
    - *(4): Interviewee 2 experiences the points of attention presneted in each section as valuable.*

Ease of use

- On a scale of 1 to 5, how user-friendly do you find this litepaper?
    - *(5): Interviewee 1 states that the litepaper is well organized and clearly written.*
    - *(4): Interviewee 2 states that the litepaper is good readable and has a clear structure.*
- On a scale of 1 to 10, how clear and understandable do you find this litepaper?

- *(5): Interviewee 1 states that for him, the litepaper contains mostly new technical information, however the litepaper sufficiently appeals to his imagination partly because of the literature and references in relation with the ethical framework, data ethics and data science supporting it.*
- *(5): Interviewee 2 states that with knowledge about software development and experience, the litepaper is understandable.*

Subjective Norm / Social Factors

- On a scale of 1 to 5, to what extent would colleagues appreciate or accept if you used this litepaper?
  - *(3): Interviewee 1 states that it is hard to say if colleagues would appreciate or accept the application of the litepaper. In the best case, data scientist already have taken aspects of the litepaper into account, but the interviewee does not have a concrete idea yet.*
  - *(3): Interviewee 2 has not given an explanation.*

Voluntariness

- On a scale of 1 to 10, to what extent are you willing to use this litepaper?
  - *(3): Interview 1 - see previous answers.*
  - *(4): Interviewee 2 has not given an explanation.*

Compatibility

- On a scale of 1 to 10, to what extent is this litepaper compatible with your field of work/profession?
  - *(1): Interviewee 1 does not deal with data science on a daily basis, but does mainly connect colleagues from various disciplines and areas of expertise in this field for implementation.*
  - *(4): Interviewee 2 states that they focus rather indirectly than directly on automated decision making systems for the insurance industry.*

7 Summary, Conclusions, Discussion and Further Research

In this research, we have followed a Design Science Research methodology approach to answer the main research question: *"What will be the most practical solution to ensure the traceability of automated outcomes in the insurance industry"*. We have discussed the current situation and identified the problem based on various cases and reports which have shown the urgency for a practical solution for traceability. There is a need for developing traceability requirements and a base set of traceability trails by standard setting bodies in cooperation with academia and the Dutch insurance industry (Brundage et al., 2020).

7.1 Summary

Having identified the current situation and the existing problems, we conducted exploratory interviews (5) to get a broader understanding about the industry specific context and elicitate the gaps, needs and objectives for a potential solution. Based on the exploratory interviews, the findings show that companies do know about traceability guidelines and requirements. However, they do not use official policy documents in order to standardize the process of complying to the traceability requirements. Most of the time, they adhere best practices as well as their own way of documenting the ML process. Explicit cases where reproduction of an outcome is needed were not encountered by the interviewees, however they do state that they are capable of reproducing outcomes when required.

The literature indicates that there is a need for clarification of guidelines, regulations and other requirements for traceability. Not only in a theoretical sense, but mostly in a practical sense. What are these requirements? What is good enough to comply to these traceability requirements? There is a need for a base set of compliance and management, next to the best practices which are already used. Finally, companies state that it would be useful to have such a practical solution, however they also say that over-regulation also could be a danger. Overall, we do think we have a good picture of the current state of guaranteeing traceability in practice.

Whilst conducting the exploratory interviews, we also conducted a literature study to gain more (technical) knowledge about traceability. Here we identified the definition of traceability and put the definition into perspective with other concepts such as reproducibility, explainablity and auditability. With the established definition, we consulted various sources such as existing and upcoming regulations, guidelines and earlier research about traceability and lineage. Lastly, we investigated which requirements apply and what the main focal points are to ensure traceability including practical examples (e.g. model cards, datasheets for datasets, reproduciblity factors).

7.2 Conclusions

We can conclude that as practical solution, a litepaper fits the best considering the elicited needs of the interviewees. In this litepaper, we have mapped the findings of the exploratory interviews and the literature study to a practical working document. This document is structured based on this thesis, which firstly elaborates on the requirements for traceability in the Dutch Insurance Industry and thereafter discusses practical focal points to realise these requirements. This litepaper is not meant as a standard or as regulatory checklist, but rather as a hands-on review on the topic of traceability, what it requires and how to facilitate traceability.

Sub questions

The answers for the sub research questions are as follows:

- *To what degree does the definition of traceability of automated systems apply in the field of Explainable AI?*

  In this research, the definition of traceability extends to documenting the development-, testing- and validation stage, as well as all the (possible) outcomes by means of recorded identification. For example data sets, processes, data gathering, data labeling and algorithms used (European Commission, 2019). This definition can't be limited because documenting these stages can be done in various ways. However, we have summarized some main factors to consider in the litepaper. These factors are also discussed in the literature study and consist of: design & development artifacts, standardized disclosures, reproducibility, monitoring, complementing data documentation and understand-ability.

- *What are the existing solutions to ensure traceability of automated outcomes?*

  We have identified several categories of solutions which could help to enable traceability. These categories are based on the assessment list of the European Commission (2019) and gives a good overview of what to document in each stage of the development of an automated decision making system. It also distinguishes between rule-based systems as well as learning-based systems. With this structure, we investigated various authors who did research on the topic of traceability in that specific category. So has Kroll (2021) identified various requirements and practical solutions in order to attain a certain level of traceability, such as documenting various design and development artifacts. In turn, these requirements fit in the framework of the European Commission (2019). Another author, Mora-Cantallops et al. (2021) reviewed various tools for attaining a level of traceability and presents various elements required for replicability. Finally, one exact solution does not exist due to the differentiation of possible methods, systems or models. However, the factors presented in 4.3 should give a good overview of what to account for when implementing traceability.

- *What is the legislation of traceability of automated outcomes?*

  There is currently no legislation that applies exactly to the traceability of automated outcomes as we have defined. However, according to AFM (2018), the Financial Supervision Act does state automated outcomes, in the form of financial advice, are treated the same as advice from a human agent. Hence it can be stated that when implementing traceability for automated systems, the same regulation applies. Upcoming is the drafted AI-act (European Commission, Directorate-General for Communications Networks, Content and Technology, 2021) which aims to introduce a regulatory and legal framework for AI. Documentation and traceability is mentioned as important requirement for high-risk AI systems, but only *Article 12 Record-Keeping* goes into detail of what to document and which logging capabilities a system should have.

Artifact validation
For the validation of the artifact, based on the answers of the questions presented in section 6.5, we can conclude the following:

- For behavioral intention, it can be concluded that the artifact will be disseminated into the organisations. However, it will be most likely that it will be disseminated to selected recipients in the organisation.

- For the usefulness of the artifact, it can be concluded that the artifact does contribute to new knowledge about traceability requirements. However, due to the nature of the job descriptions of the interviewees, it can not be fully said that the litepaper will make it easier to do their daily work. One company says that they already perform many steps in their processes to ensure traceability, but in terms of documentation regarding traceability , they could still make improvements. For it concrete usefulness, it can be concluded that the interviewees find it medium-high useful.

- For the ease of use, it can be concluded that the litepaper is user friendly, clear and understandable, e.g: ”It is well organized and clearly written”.

- For the social factors, it can be concluded that it is hard to say if colleagues appreciate or accept the use of the litepaper. This is because one company did not respond to the question and the other interviewee does not have a concrete idea (yet).

- For voluntariness (the extend that the interviewees are willing to use the litepaper), an interviewee referred to ”the previous answers”. Here the average score is medium-high, hence in combination with the usefulness determinant, we will conclude that the interviewees are willing to use the litepaper.

- For the compatibility (the extent that the litepaper is compatible with the interviewees profession), one interviewee responded that as ethicist not directly deal with data science on a daily basis, but mainly connect colleagues from various disciplines. Another interviewee also state that they rather focus indirectly on automated decision systems. With also the separate scores, we cannot draw a decisive conclusion.

7.3 Discussion, Further Research and Takeaways

We came across a few limitations in this research. Firstly, we could not exactly pinpoint the scope of traceability due to the broad range of automated decision making systems. At the moment of writing, the exact traceability requirements differ in each different automated decision making system. This makes enabling traceability for e.g. traditional software systems easier rather than e.g. a deep neural network. This limitation led to the obligation to stay general in terms of drafting traceability requirements. For further research and as takeaway, our recommendation is to research one specific automated decision making system and identify the traceability requirements only for this ADM. Here by, you can better target the specific needs and compile more detailed and restricted practical requirements.

Second, at the problem identification stage (exploratory interviews), we conducted five interviews at four different companies. In terms of representability, the amount of conducted interviews are low and the findings could differ when interviewing other companies. However, we have interviewed four different kinds of companies: a major insurance group, a smaller insurance intermediate, a software vendor and a major insurance broker. Also, note that we left the major insurance broker out of our data. This is because we have recieved this data at the final stage of our research. However, we retrieved valuable insights and knowledge, which was in line with the findings of the other interviews. For further research, we recommend to interview more companies and more different positions, but as already stated, in a narrower scope such as one sort of automated decision making system.

Third, a limitation we came across was the type of practical solution as stated in our research question. At first, the practical solution was aimed at a more hands on solution, such as a standard or a type of versioning system for traceability. However, the scope and depth of complexity of the topic restricted this kind of solution. Nonetheless, for further research we can recommend a few steps in the direction of this kind of practical solution. We recommend to make smaller traceability standards for the each type of requirement and each type of ADM. For example, a standard for the logging capabilities of specifically rule based systems, or a standard for what to document and save to reproduce an outcome based on a supervised regression model. We have presented the beginning of what this standard should entail, which can be found in the "requirements and focal points to ensure traceability" section of this study. Another step we would recommend is to focus on what is known in terms of regulation, guidelines and obligations. For example, there is an upcoming AI act drafted, but everything in it is not yet definitive. Other acts, e.g. as the Financial Supervision Act, are as of now definitive and standard should be based on complying to such acts.

With this study, we hope we have contributed to the practice of enabling traceability for ADM's. With being able to trace back outcomes, companies can shield their selves from costly juridical procedures and account for their decisions (e.g. HDN case). This not only serve the insurance companies, but also the consumer who can understand how and why their insurances are established.

We have showed why traceability is an urgent topic in the Dutch Insurance Industry, what the current needs and gaps are, which requirements and focal points fit these needs and eventually presented an artifact containing a hands on solution. However, the topic of traceability and AI, as well as its surrounding regulations, is very volatile. In a few years, the upcoming AI act applies and could be different from this research. This results in the limitation that the artifact is subject to change, because it is a snapshot which should be viewed as advisory rather than obligatory. Nevertheless, with this artifact we hope we have contributed to the advancing practice of traceable outcomes in the Dutch Insurance Industry.

References

About ml reference document, research themes on documentation for transparency. (2021, October 7). Retrieved October 10, 2022, from hthttps://partnershiponai.org/paper/about-ml-reference-document/6/#Section-2-3-1

AFM. (2018). *Visie op roboadvies kansen, zorgplicht en aandachtspunten.* https://www.afm.nl/~/profmedia/files/onderwerpen/roboadvies-sav/visie-roboadvies.pdf

AFM & DNB. (2019). *Artificiële intelligentie in de verzekeringssector een verkenning.*

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., & Varshney, K. R. (2018). Factsheets: Increasing trust in ai services through supplier's declarations of conformity. http://arxiv.org/abs/1808.07261

Barredo, A., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bertuzzi, L. . (2022). Last-minute changes to eu council's ai act text ahead of general approach. https://www.euractiv.com/section/digital/news/last-minute-changes-to-eu-councils-ai-act-text-ahead-of-general-approach/

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., . . . Anderljung, M. (2020). Toward trustworthy ai development: Mechanisms for supporting verifiable claims. http://arxiv.org/abs/2004.07213

Cambridge Dictionary. (2022). audit definition: 1. to make an official examination of the accounts of a business and produce a report 2. to go to a. . . . Learn more. https://dictionary.cambridge.org/dictionary/english/audit

De Stichting Vervoeradres. (2022). Hoe ziet de verzekeringsmarkt eruit? — SVA. https://www.sva.nl/themas/verzekeringen/praktijkboek-over-verzekeringen-de-logistiek/hoe-ziet-de-verzekeringsmarkt-eruit

Ettema, L. (2021a, March 19). *Brondata betrouwbaar of niet?* Retrieved September 26, 2022, from https://magazines.infinance.nl/infinance-magazine-01-2021/brondata-betrouwbaar-of-niet

Ettema, L. (2021b, March 26). *Ibl-rekentool geeft maandenlang verkeerd toetsinkomen af.* Retrieved September 26, 2022, from https://magazines.infinance.nl/infinance-magazine-01-2021/ibl-rekentool-geeft-maandenlang-verkeerd-toetsinkomen-af

European Commission. (2019). *High-level expert group on artificial intelligence set up by the european commission ethics guidelines for trustworthy ai.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission (Ed.). (2022, October 27). *Regulatory framework proposal on artificial intelligence.* Retrieved November 15, 2022, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Proposal Artifical Intelligence Act (2021, April). Retrieved October 20, 2022, from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2018). Datasheets for datasets. http://arxiv.org/abs/1803.09010

Gundersen, O. E., Gil, Y., & Aha, D. W. (2018). On reproducible ai: Towards reproducible research, open science, and digital scholarship in ai publications.

Gundersen, O. E., & Kjensmo, S. (2017). *State of the art: Reproducibility in artificial intelligence.* https://ojs.aaai.org/index.php/AAAI/article/view/11503

HDN. (2022, September 26). *Inkomensbepaling loondienst.* https://www.hdn.nl/inkomensbepalingloondienst/

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: Amee guide no. 131. *Medical Teacher*, *42*, 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Koster, O., Kosman, R., Mosterd, W., & Visser, J. (2021). *Explainable artificial intelligence: A checklist for the insurance market.* https://theses.liacs.nl/2129

Kroll, J. A. (2021). Outlining traceability: A principle for operationalizing accountability in computing systems. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 758–771. https://doi.org/10.1145/3442188.3445937

Leijnen, S., Kuiper, O., & van der Berg, M. (2022). *Impact your future xai in the financial sector a conceptual framework for explainable ai (xai).* https://www.hu.nl/onderzoek/projecten/uitlegbare-ai-in-de-financiele-sector

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M. A. (2021). Traceability for trustworthy ai: A review of models and tools. *Big Data and Cognitive Computing*, *5*. https://doi.org/10.3390/bdcc5020020

Nen-iso 8402:1994 nl. (1994, October). Retrieved January 5, 2023, from https://www.nen.nl/en/nen-iso-8402-1994-nl-6740

Olorisade, B. K., Brereton, P., & Andras, P. (2017). *Reproducibility in machine learning-based studies: An example of text mining.* https://doi.org/https://doi.org/10.1016/j.jbi.2017.07.010

Olsen, P., & Borit, M. (2013). *How to define traceability.* https://doi.org/10.1016/j.tifs.2012.10.003

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Raji, I. D., & Yang, J. (2019). About ml: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. http://arxiv.org/abs/1912.06166

Riemenschneider, C. K., Hardgrave, B. C., & Davis, F. D. (2002). Explaining software developer acceptance of methodologies: A comparison of five theoretical models. *IEEE transactions on Software Engineering*, *28*(12), 1135–1145.

SIVI, S. (2022, April 25). *Sivi.* Retrieved October 4, 2022, from https://www.sivi.org/over-sivi/

Verbond van Verzekeraars. (2015). *Ethisch kader datatoepassingen.* Retrieved October 20, 2022, from https://www.verzekeraars.nl/branche/zelfreguleringsoverzicht-digiwijzer/ethisch-kader-datatoepassingen

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. In *Design science research. cases* (pp. 1–13). Springer.

8 Appendix A: Litepaper Traceability

Herleidbaarheid van de resultaten van Onbemenste Toepassingen

Universiteit Leiden

Bij ons leer je de wereld kennen

Herleidbaarheid van de resultaten van Onbemenste Toepassingen

Een litepaper over de aandachtspunten bij het borgen van herleidbaarheid van resultaten van onbemenste toepassingen in de Nederlandse verzekeringsindustrie.

Auteur
Kas Bekker

Afdeling
Stichting SIVI

Datum
12-Dec-22

Project type
Litepaper

Versie
1.2

Samenvatting

Om betrouwbare en transparante dienstverlening aan verzekeringsklanten te bieden, dienen adviezen en beslissingen die geheel of gedeeltelijk afkomstig zijn van een geautomatiseerd proces herleidbaar te zijn. Dit houdt in dat door middel van geregistreerde documentatie, de totstandkoming van een resultaat of beslissing inzichtelijk kan worden gemaakt.

Omdat herleidbaarheid en haar eisen vaak nog een abstract begrip is, biedt dit document verduidelijking van de eisen als aandachtspunten om herleidbaarheid te waarborgen. Deze aandachtspunten zijn gestructureerd op basis van bestaande richtlijnen en eisen, waarna een totaalbeeld ontstaat van mogelijke stappen die de herleidbaarheid van resultaten kunnen ondersteunen.

Deze litepaper is bedoeld voor zowel ontwikkelaars als beleidsmakers die zich inzetten om hun onbemenste toepassingen te laten voldoen aan wettelijke en ethische normen en eisen. Allereerst zullen de relevante eisen en richtlijnen rondom herleidbaarheid worden uiteengezet. Vervolgens presenteren we de bijbehorende aandachtspunten via de volgende categorieën: Design & Ontwikkeling, Testen & Validatie, Data & Uitkomsten.

Inhoudsopgave

1. Inleiding

Het gebruik van onbemenste toepassingen (OT) neemt toe in de Nederlandse verzekeringsindustrie. Volgens de Autoriteit Financiële Markten (AFM) en de Nederlandsche Bank (DNB) brengen OT niet alleen nieuwe mogelijkheden, maar ook bijkomende onzekerheden en risico's. Adequate aandacht, bewustzijn en maatregelen zijn dan ook nodig om deze risico's en onzekerheden te mitigeren. Het wordt steeds belangrijker om transparant te zijn met zowel de inzet van OT als de uitkomsten en besluiten die OT opleveren. Met transparantie wordt bedoeld het inzichtelijk maken van de totstandkoming van een resultaat. Hierin spelen twee aspecten een belangrijke rol, namelijk herleidbaarheid en uitlegbaarheid.

Dit document gaat in op de **herleidbaarheid** (traceerbaarheid) van onbemenste toepassingen. Herleidbaarheid heeft betrekking op de mogelijkheid om data, ontwikkelings- en inzet processen van een geautomatiseerd systeem bij te houden door middel van gedocumenteerde geregistreerde identificatie. In samenwerking met stichting SIVI en Universiteit Leiden, brengen wij verschillende punten onder de aandacht die helpen om de herleidbaarheid van uitkomsten te kunnen waarborgen.

Om een zo goed mogelijk overzicht te geven, is dit document gestructureerd in de volgende onderdelen:

- Eisen en richtlijnen rond herleidbaarheid
- Aandachtspunten voor het borgen van herleidbaarheid
- Referentielijst
- Appendix A: Begrippenlijst

We verduidelijken de eisen en richtlijnen rond herleidbaarheid samen met de bijbehorende bronnen. Vervolgens zullen we verschillende punten onder de aandacht brengen die passen bij het borgen van herleidbaarheid in het kader van deze eisen en richtlijnen.

Tot slot is dit document bedoeld voor iedereen die betrokken is bij de inzet van een onbemenste toepassing. Dit kan een ontwikkelaar zijn, maar ook beleidsmakers, ethici of (project) managers.

2. Eisen en richtlijnen rond herleidbaarheid

Er is veel te doen om de eisen en richtlijnen bij het ethisch verantwoord inzetten van onbemenste toepassingen. Er zijn (nog) geen officiële wetten die stellen wat er gedocumenteerd moet worden. Echter zijn er wel toepasselijke bronnen voor de inzet van onbemenste toepassingen. Zo is door het Verbond van Verzekeraars, specifiek voor de verzekeringsindustrie, een Ethisch Kader gepresenteerd. Naast dit Kader bestaan er Europese richtlijnen, gepresenteerd in Ethical Guidelines for Trustworthy AI. Ook bespreken we een visie van de AFM op het "Robo Advies" die ingaat op hoe de wettelijke zorgplicht zich verhoudt tot adviezen die afkomstig zijn van (deels) geautomatiseerde systemen. Als laatste bespreken we kort kernpunten uit relevante vakliteratuur over herleidbaarheid. Ten behoeve van verduidelijking van deze eisen en richtlijnen rondom herleidbaarheid, hebben we de belangrijkste punten uit bovenstaande bronnen uiteengezet. Let op! Deze eisen en richtlijnen zijn niet sluitend en zijn ervoor bedoeld om verduidelijking te geven voor wat er nodig om herleidbaarheid te borgen. Deze eisen en richtlijnen zijn niet bedoeld als juridisch of wettelijk minimum.

2.1 Herleidbaarheid en het Ethisch Kader van het Verbond van Verzekeraars

Het Ethisch Kader van het Verbond van Verzekeraars vormt een instrument voor zelfregulering bij de inzet van (gedeeltelijke) onbemenste toepassingen. Dit kader is een 'principle based' kader wat zich richt op het borgen van verantwoordelijke en vertrouwenwekkende toepassing van data. Sinds 01-01-2021 geldt dit Kader als zelfregulering. Vanaf begin 2023 toetst Stichting Toetsing Verzekeraars (Stv) het Ethisch Kader op naleving. Deze toetsing wordt o.a. gedaan door het aanleveren en controleren van verschillende documenten en het houden van gesprekken over de borging van het Ethisch Kader. Meer informatie over de toetsing van het Ethisch Kader is hier te vinden.

De punten die voor herleidbaarheid relevant zijn, staan hieronder kort samengevat:

- Stichting Toetsing Verzekeraars vraagt bij een audit/toetsing om een document die toelicht op de wijze hoe onbemenste toepassingen gemonitord worden.
- In de gesprekken zal Stichting Toetsing Verzekeraars de volgende sleutelvragen stellen die relevant zijn voor herleidbaarheid:
    - Kan er beroep worden gedaan op menselijke tussenkomst en wat gebeurt er dan?
    - Wat is de uitleg aan klanten over de uitkomsten van datagedreven toepassingen?

Het Ethisch kader omvat verder 30 normen die niet alleen gelden voor AI, maar voor alle moderne datagedreven besluitvorming die van invloed is op het klant vertrouwen. Als we kijken naar herleidbaarheid in het Ethisch Kader, zijn de volgende normen belangrijk om in acht te nemen als bron van eisen voor herleidbaarheid.

Tabel 1: Aandachtspunten uit het Ethisch Kader relevant voor Herleidbaarheid.

Aandachtspunt	Norm voor verzekeraars	Toelichting, zie verder Toolkit Ethisch Kader
Betrouwbaarheid en reproduceerbaarheid	6. Verzekeraars monitoren of gebruikte data gedreven systemen in overeenstemming met vooraf gestelde doelen, doelstellingen en beoogde toepassingen werken	Het monitoren van onbemenste toepassingen kan op verschillende methoden. Wat het meest wordt teruggezien zijn: logging, trend rapportages, performance rapportages, real-time monitoring, geautomatiseerde controles en validaties.
Transparantie	17. Voordat wij data gedreven systemen inzetten bedenken wij hoe we zo goed mogelijk uitleg kunnen geven aan klanten over de uitkomsten van de toepassing.	In hoeverre is er in de praktijk getest/gecontroleerd hoe de uitkomsten van het systeem worden geïnterpreteerd door de belanghebbenden (klanten, medewerkers, etc.) ?

Controleerbaarheid	23. Wij zorgen voor een intern controle- en verantwoordingsmechanisme voor het gebruik van AI systemen en de gebruikte databronnen.	Hoe is vastgelegd voor de toepassing, a) wat het doel is, b) wie de eigenaar is, c) welke databronnen zijn gebruikt voor de ontwikkeling van de toepassing, d) welke databronnen worden gebruikt voor de inzet van de toepassing, e) wat de relevante risico's zijn bij de inzet van de toepassing

2.2 Herleidbaarheid in "Ethical Guidelines For Trustworthy AI."

Het Ethisch Kader van het Verbond van Verzekeraars is gebaseerd op aanbevelingen van de High-Level Expert Group on Artificial Intelligence (AI HLEG, 2019)[1]. Dit adviesorgaan van de Europese Commissie heeft 7 vereisten vastgesteld om ethische en verantwoordelijke AI te waarborgen: 1) Menselijke autonomie en controle, 2) Technische robuustheid en veiligheid, 3) Privacy en Data Governance, 4) Transparantie, 5) Diversiteit, Non-discriminatie en Rechtvaardigheid, 6) Maatschappelijk welzijn, 7) Verantwoording.

Volgens de Europese Commissie AI High Level Expert Group (AI HLEG, 2019) bestaat vereiste 4) transparantie uit 3 componenten: **herleidbaarheid**, uitlegbaarheid en communicatie. Voor herleidbaarheid heeft de AI HLEG een assessment list opgesteld met punten die in acht genomen kunnen worden om herleidbaarheid te waarborgen.

Herleidbaarheid assessment list. (AI-HLEG, 2019)
Hebt u maatregelen ingesteld waarmee de herleidbaarheid kan worden gewaarborgd? Daarbij kan het gaan om de documentatie van:

De voor het **ontwerp en de ontwikkeling** van het algoritmische systeem gebruikte methoden:

- in geval van een AI-systeem op basis van regels moet de programmeermethode of de manier waarop het model is gebouwd, worden gedocumenteerd;
- in geval van een AI-systeem op basis van leren, moet de trainingsmethode van het algoritme, met inbegrip van welke input data er is verzameld en geselecteerd en de manier waarop dit is gebeurd, worden gedocumenteerd.

De voor het **testen en valideren van** het algoritmische systeem gebruikte methoden:

- in geval van een AI-systeem op basis van regels, moeten de voor het testen en valideren gebruikte scenario's of situaties worden gedocumenteerd;
- in geval van een model op basis van leren, moet de voor het testen en valideren gebruikte informatie worden gedocumenteerd.

De **resultaten** van het algoritmische systeem:

- de resultaten of de door het algoritme genomen beslissingen, alsook potentiële andere beslissingen die in andere situaties zouden ontstaan (bijv. voor andere subgroepen of gebruikers) moeten worden gedocumenteerd.

Figuur 1: Assessement list for Traceabiltiy.

[1] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

2.3 Herleidbaarheid en de Wet Financieel Toezicht.

Naast de richtlijnen voor "Trustworthy AI" vanuit de Europese Unie zijn er relevante juridische aandachtspunten bij herleidbaarheid. De AFM gaat in de publicatie "Visie op Roboadvies"[2] verder met deze aandachtspunten. Echter, "Robo Advies" is een verouderde term. In een Wijzigingsbesluit Financiële Markten 2021[3] wordt de term "geautomatiseerd advies" gebruikt. Om actueel te blijven vervangen wij dan ook de term "Robo Advies" voor "geautomatiseerd advies". Bij het borgen van herleidbaarheid, zijn volgende punten uit deze publicatie belangrijk om in het achterhoofd te houden:

- **Volgens de zorgplicht wordt er geen onderscheid gemaakt tussen een fysiek advies of geautomatiseerd advies .**
  Een advies dat tot stand is gekomen op basis van een resultaat van een onbemenste toepassing wordt ook wel een geautomatiseerd advies genoemd. Volgens de AFM wordt in de Wet Financieel Toezicht geen onderscheid gemaakt tussen een geautomatiseerd advies en een fysiek advies. Het uitgangspunt is dat een geautomatiseerd advies aan dezelfde invulling van de zorgplicht (artikel 4:20, 4:23 en 4:90 Wft) moet voldoen als bij een fysiek advies.

- **Bij uitbesteding is de aanbieder van het advies altijd (eind) verantwoordelijk.**[4]
  Wanneer er sprake is van uitbesteding, is de aanbieder van een advies nog steeds verantwoordelijk bij de uitbesteding. Er dient een proces ingericht te zijn waarmee de kwaliteit van de dienstverlening van een leverancier gemonitord kan worden. De aanbieder is dus ook verantwoordelijk voor de kwaliteit van een geautomatiseerd advies, als de (functionele) beheersing van de onbemenste toepassing. Wanneer gesproken wordt over de eisen van herleidbaarheid, is het dus aan de aanbieder om functies aan te vragen die het mogelijk maken om aan de eisen te kunnen voldoen.

- **Een geautomatiseerd advies moet herleidbaar en reproduceerbaar zijn.**
  Een **g**eautomatiseerd advies afkomstig van een onbemenste toepassing moet een gedegen audittrail hebben. In andere woorden, bij een geval van een audit moet een uitkomst voldoende herleidbaar en reproduceerbaar zijn. Hierin is goede vastlegging via gestructureerde documentatie essentieel. Deze vastlegging maakt het mogelijk dat een klant, toezichthouder of eventueel (vervolg) adviseur kan nagaan hoe een advies tot stand is gekomen. Niet alleen keuzes in het adviesproces dienen vastgelegd te worden, maar ook toekomstige wijzigingen moeten aanvullend worden opgeslagen.

2.4 Herleidbaarheid vereisten uit de vakliteratuur

Naast de richtlijnen van de Europese Commissie en het Ethisch Kader, zijn er verschillende onderzoeken geweest naar het operationaliseren van herleidbaarheid. Zo heeft Kroll[5], o.a. op basis van de richtlijnen van de Europese Commissie, vijf categorieën van vereisten geïdentificeerd waar rekening mee moet gehouden worden om een bepaalde mate van herleidbaarheid te realiseren. Deze vijf vereisten omvatten onder andere het rekening houden met:

1.	Ontwerp transparantie	*het vrijgeven van ontwikkel en ontwerpdocumenten, data en code*
2.	Reproduceerbaarheid	*het kunnen reproduceren van een uitkomst*
3.	"Operational Recordkeeping"	*het monitoren en bijhouden van gestructureerde "logs"*
4.	Begrijpelijkheid	*het begrijpelijk maken van de vrijgegeven documentatie*
5.	"Auditability"	*de mate waarop een OT controleerbaar is door een externe partij*

[2] https://www.afm.nl/~/profmedia/files/onderwerpen/roboadvies-sav/visie-roboadvies.pdf

[3] Wijzigingsbesluit Financiële Markten 2021 artikel 32 da, https://www.rijksoverheid.nl/documenten/kamerstukken/2021/11/24/bijlage-1-wijzigingsbesluit-financiele-markten-2021

[4] Zie 4.1.3 in https://www.afm.nl/~/profmedia/files/onderwerpen/roboadvies-sav/visie-roboadvies.pdf

[5] Kroll, J. A. (2021, March). Outlining traceability: A principle for operationalizing accountability in computing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 758-771). https://doi.org/10.1145/3442188.3445937

Ook heeft Mora-Cantallops et al.[6], ook o.a. op basis van de richtlijnen van de Europese Commissie, onderzocht wat relevante tools zijn om herleidbaarheid te borgen. Naast deze tools, identificeert Mora-Cantallops et al. een minimaal beschrijvingsprofiel met verschillende elementen om aan *repliceerbaarheid* te kunnen voldoen.

Als laatste heeft Raji et al.[7] een end-to-end framework voor interne algoritmische auditing voorgesteld. Hierin wordt gesteld dat het steeds moeilijker wordt om opkomende problemen bij (AI) systemen te kunnen herleiden. Om deze "verantwoordingskloof" te verkleinen gebruikt Raji et al. verschillende fases van het ontwerp en ontwikkelproces van AI-modellen om verschillende sets van documenten en artefacten te structureren.

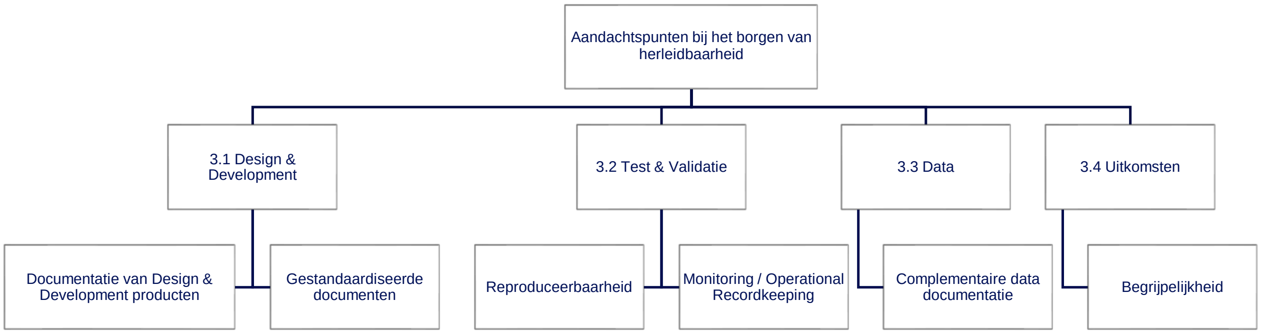
Voor dit document, gebruiken wij verschillende elementen uit bovenstaande vakliteratuur om de aandachtspunten te onderbouwen. Graag verwijzen wij naar deze auteurs voor uitgebreidere informatie.

[6] Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M. A. (2021). Traceability for trustworthy ai: A review of models and tools. *Big Data and Cognitive Computing*, *5*(2), 20. https://doi.org/10.3390/bdcc5020020

[7] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44). https://dl.acm.org/doi/abs/10.1145/3351095.3372873

3. Aandachtspunten bij het borgen van herleidbaarheid

Met het oog op de toepasbaarheid van dit document, gebruiken wij de evaluatielijst (figuur 1) afkomstig van de High-Level Expert Group on AI om de aandachtspunten voor het borgen van herleidbaarheid te structureren.


Aandachtspunten bij het borgen van herleidbaarheid
3.1 Design & Development
3.2 Test & Validatie
3.3 Data
3.4 Uitkomsten
Documentatie van Design & Development producten
Gestandaardiseerde documenten
Reproduceerbaarheid
Monitoring / Operational Recordkeeping
Complementaire data documentatie
Begrijpelijkheid

Figuur 2: De structuur van de aandachtspunten voor herleidbaarheid in praktijk.

Deze assessment lijst maakt expliciet onderscheid tussen OT die gebaseerd zijn op regels of gebaseerd zijn op leren. Aangezien wij ook rekening houden met onbemenste toepassingen die niet per se een AI systeem zijn, gelden deze aandachtspunten ook voor OT die buiten de categorie van AI vallen.

Om de aandachtspunten te onderbouwen, hebben wij onder elk aandachtspunt in een tabel ondersteunende vragen gezet die het doel hebben de lezer aan het denken te zetten over of hij/zij bepaalde maatregelen heeft ingesteld, en op welke manier. Naast deze vragen staat de toelichting en/of voorbeelden die deze vragen onderbouwen.

3.1 Design & Ontwikkeling

Herleidbaarheid vereist welke bepaalde ontwerpkeuzes van een onbemenste toepassing zijn gemaakt en waarom. In het geval van een model op basis van regels (rule-based), moet de programmeermethode of de manier waar op het model is gebouwd, worden gedocumenteerd. In het geval op een model op basis van leren (learning-based), moet de trainingsmethode van het algoritme, met inbegrip van welke input data er is verzameld en geselecteerd en de manier waarop dit is gebeurd, worden gedocumenteerd. Bij de fase van design & ontwikkeling willen wij 2 punten onder de aandacht brengen die deze documentatie ondersteunen.

3.1.1 Aandachtspunt 1: Documentatie van Design & Development producten

Een primaire eis van herleidbaarheid is dat keuzes die door systeemontwerpers worden gemaakt, inzichtelijk worden gemaakt aan de belanghebbenden van een onbemenste toepassing. Dit wordt in het algemeen bereikt door het beschikbaar stellen van code en data, met eventueel nog extra documentatie. Het eerste aandachtspunt is dan ook het beschikbaar stellen van alle documenten die samengesteld zijn bij de ontwikkeling en het ontwerp van een onbemenste toepassing. Denk hierbij aan code, afleidingsregels, modelkeuzes, hyperparameters etc.

Een veelgebruikte methode bij het ontwikkelen en ontwerpen van een onbemenste toepassing, is het vaststellen van een requirements document. Requirements geven weer hoe een doel van een bepaald systeem is benaderd door de ontwerpers en hoe dit doel gerealiseerd kan worden. Voldoende herleidbaarheid houdt dan ook in dat keuzes en parameters achter het ontwerp van een onbemenste toepassing kenbaar worden gemaakt. Het vrijgeven van documentatie, zoals een requirements document, achter de keuzes en parameters dragen dan ook bij aan herleidbaarheid. (Kroll, 2021)

Aandachtspunten voor de documentatie van Design & Development producten.

Aandachtspunten	**Toelichting**
Aandachtspunt 1: Documentatie van Design & Development producten	
Welke code en/of gebruikte datasets worden bij opvraag[8] inzichtelijk en eventueel beschikbaar gemaakt?	*Een primaire eis van herleidbaarheid is dat keuzes die door ontwerpers zijn gemaakt, inzichtelijk worden gemaakt aan belanghebbenden van een onbemenste toepassing. Het beschikbaar stellen van code en data kan dit inzicht geven.*
Hoe wordt het ontwerp (design) van een onbemenste toepassing gedocumenteerd?	*Hoe een ontwerp van een OT gedocumenteerd wordt, bepaalt de mate van herleidbaarheid bij een uitkomst en draagt bij aan het inzichtelijk maken van ontwerpkeuzes. Denk aan: requirements documenten, use cases, afleidingsregels etc.*
Hoe wordt het testproces van een onbemenste toepassing gedocumenteerd? Met welke situaties en scenario's is er getest?	*Naast de documentatie van het design, is het van belang dat de testprocedure is vastgelegd en opgevraagd kan worden. Denk aan de bijbehorende (test) scenario's, gemaakte (test) keuzes etc.*
Bij learning-based modellen, hoe documenteren jullie de totstandkoming van het model?	*Specifiek bij learning-based modellen is het van belang dat de trainingsmethode, met inbegrip welke input data er is verzameld en geselecteerd, wordt vastgelegd om herleidbaarheid te borgen.*

[8] Bij de opvraag van de totstandkoming van een uitkomst door bijv. een klant, belanghebbende of jurist.

3.1.2 Aandachtspunt 2: Gestandaardiseerde eenduidige (uitleggende) documenten.

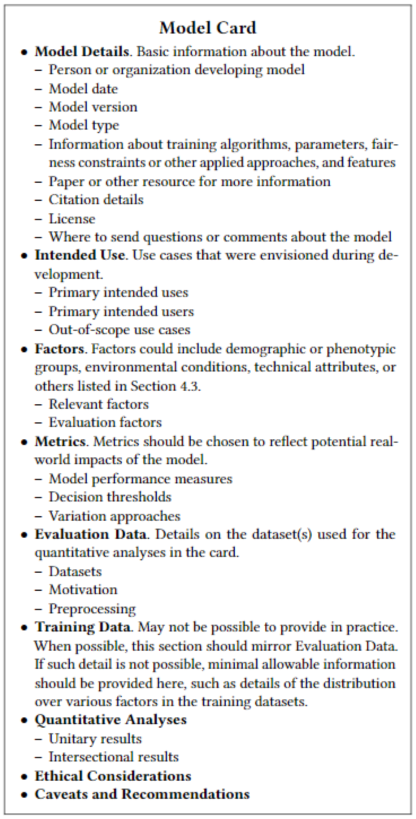
Requirements zijn niet altijd formeel gespecificeerd of kunnen vanwege privacy en/of copyright redenen niet direct worden vrijgegeven. Het tweede aandachtspunt gaat dan ook om het documenteren van een model en haar uitkomsten doormiddel van eenduidige (disclosed) gestandaardiseerde documenten.

Ook in het geval van rule-based systemen kan het zijn dat vanwege ondoorzichtigheid door grote hoeveelheden regels, de regels niet meer goed begrepen worden. Dit zorgt voor onduidelijkheden over wat de gevolgen zijn van een specifieke wijziging in een deel van een toepassing.

Tot slot zijn veel toepassingen het resultaat van trial en error (bijv. unsupervised learning). Deze systemen hebben niet per se een vooraf gesteld ontwerpdocument klaar staan en documentatie komt in deze gevallen meestal tot stand "on the fly".

Om problemen zoals hierboven beschreven aan te pakken en alsnog de herleidbaarheid te borgen, is het belangrijk om aandacht te schenken aan documentatie waarmee modellen en uitkomsten in één oogopslag samengevat kunnen worden. Artefacten die hierbij kunnen helpen zijn bijvoorbeeld.

- ***Design Checklist***. Een design checklist is een methode om alle verwachte en gegenereerde documentatie te inventariseren. Doormiddel van het vaststellen van zo'n checklist, kan worden afgegaan welke documenten gebruikt zijn in de design fase. Bij de ontwikkeling is het al belangrijk dat z'n design checklist al vastgesteld wordt, waarna hier later altijd nog op terug kan worden gevallen (Raji et al, 2020).
- ***Model Card***. Een recente standaard om meer controleerbare en generiekere documentatie te genereren is die van een Model Card, ontwikkeld door Mitchel et al. Een model card bevat informatie hoe een model is gebouwd, de gekozen aannames, het type gedrag alsook de gemaakte design keuzes. Een robuuste model card is de sleutel tot het documenteren van de intenties van het model, alsook informatie over de totstandkoming van resultaten.



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Aandachtsputnen voor gestandaardiseerde eenduidige (uitleggende) documenten.

Aandachtspunten	Toelichting
Aandachtspunt 2: Gestandaardiseerde eenduidige (uitleggende) documenten	
Welke methoden van documentatie gebruiken jullie om een onbemenste toepassing in één gestandaardiseerd document vast te leggen?	*Eenduidige gestandaardiseerde documenten dragen bij aan herleidbaarheid omdat deze documenten in één oogopslag overzicht en inzicht geven aan de gehele onbemenste toepassing, ook in het geval dat volledige herleidbaarheid niet te borgen valt. Denk hier aan: model cards, datasheets for data sets of eigen methoden.*

3.2 Testen & Validatie

Bij het testen en valideren van AI-modellen, vereist herleidbaarheid dat de gebruikte situaties en scenario's als de gebruikte data worden gedocumenteerd. Herleidbare systemen moeten informatie vrijgeven over hun test- en evaluatieplannen. De reden hiervoor is dat herleidbaarheid vereist dat ontwikkelaars en de belanghebbenden de kloof tussen wat bekend is bij de ontwikkelaars (ontwerpbeslissingen, test en evaluatieregime) en wat bekend is bij de belanghebbenden (die zien niet het resultaat van alle ontwikkelingstests) minimaliseren (Kroll, 2021). Om deze kloof te minimaliseren, en validatie mogelijk te maken stellen wij reproduceerbaarheid en operationele recordkeeping als relevante aandachtspunten om herleidbaarheid te borgen.

3.2.1 Aandachtspunt 3: Reproduceerbaarheid

Om herleidbaarheid te borgen, moet het mogelijk zijn om uitkomsten afkomstig van een onbemenste toepassing te kunnen reproduceren. Met het reproduceren van een uitkomst, kan een externe belanghebbende verifiëren waarom en hoe een systeem is ontworpen en op welke manier.

Als we het hebben over reproduceerbaarheid, bedoelen we de mate van reproduceerbaarheid wat ook *repliceerbaarheid* wordt genoemd. Dit houdt in dat een onafhankelijke partij (niet de oorspronkelijke "maker" van de onbemenste toepassing) een uitkomst kan repliceren met gebruik van de artefacten van de oorspronkelijke "maker". Denk hierbij aan de gebruikte code, dataset, software, hardware etc. Met het kunnen repliceren/ reproduceren van een uitkomst, kan een uitkomst gevalideerd worden door een onafhankelijke partij.

Concept en model drift

Een bekend risico bij het reproduceren van een uitkomst is het verlies of verandering van de juiste ontwikkelings-informatie en/of omgeving. Denk hier aan de historische code, data, model of andere componenten die gebruikt zijn op een specifiek tijdstip voor een bepaalde uitkomst of advies.

Wanneer een omgeving van een onbemenste toepassing verandert en de historische data die gebruikt was niet meer representatief is voor de nieuwe omgeving/situatie, is er sprake van "concept drift". Wanneer sprake is van een veranderde input data, en deze input wordt gebruikt voor een nieuwe versie van een onbemenste toepassing, spreekt men van model drift. Bij het reproduceren van een uitkomst, moet rekening gehouden met beide vormen van *drift*. Deze vormen van *drift* moeten dan ook tijdig gesignaleerd worden, waarna er gestart kan worden met het verbeteren van een toepassing. (KPMG, 2020)

Reproduceerbaarheid factoren

Olorisade, Brereton, & Andras (2017) hebben in een onderzoek naar de reproduceerbaarheid van Machine Learning gebaseerde studies een set aan factoren geïdentificeerd dat reproduceerbaarheid kunnen faciliteren. Ongeacht welke tooling of omgevingen gebruikt worden, kan er mimimaal rekening gehouden worden met:

Tabel 4: Reproduceerbaarheid factoren.

Factoren	Vragen die gesteld kunnen worden	Toelichting
Dataset	Is de gebruikte (ruwe) dataset opgeslagen op een bereikbaar medium? Is deze dataset geleverd met een toelichting, bijv. een *Datasheet?*	- Informatie over de locatie alsook het verkrijgingsproces van een dataset is vereist voor reproductie.
Preprocessing	Kan een onafhankelijke partij de (ruwe) dataset (opnieuw) preprocessen naar de dataset die gebruikt wordt voor het desbetreffende algoritme?	- Het proces van het preprocessen van input data door het labelen, noise reductie als ook het transformeren naar een acceptabele input.
Data partities	Zijn er details over hoe een dataset opgedeeld is in test, training en validatie set?	- Details over de training, test en eventuele validatie dataset is vereist voor reproductie.
Model training	Is er informatie beschikbaar die elke beslissing bij het trainen van het model rechtvaardigt?	- Details over hyperparameters[9], technieken, keuzes - Beschikbare code, algoritmes etc.
Model assessment	Is er een performance measure (prestatie maat) genomen om de prestaties van het model te meten?	- Details over de keuze over de performance measure en voorgestelde technieken.
Randomization Control	Wordt er gebruik gemaakt van randomisatie? Zo ja, hoe leg je dit vast voor reproductie?	- Details over het randomisatie proces, gebruikte "seeds[10]" en gemaakte keuzes.
Software environment	Kan de softwareomgeving waarin het model tot stand is gekomen gereproduceerd worden?	- Details over gebruikte packages, modules, versienummers, OS versies en andere dependencies.
Hardware environment	Kan de uitkomst enkel gereproduceerd worden op een specifieke hardware omgeving?	- Indien nodig, details over de vereiste hardware. Denk aan opslagruimte, werkgeheugen (RAM), GPU-kracht etc.

[9] Bij Machine Learning, een "hyperparameter" die invloed heeft op het leerproces en wordt gebruikt om het leerproces te controleren.

[10] Een "seed" is een pseudorandom waarde die gebruikt wordt om een random nummer te genereren.

Aandachtspunten voor reproduceerbaarheid.

Aandachtspunten	**Toelichting**
Aandachtspunt 3: Reproduceerbaarheid	
Hoe faciliteren jullie de reproduceerbaarheid van een uitkomst?	*Reproduceerbaarheid draagt bij aan de herleidbaarheid omdat de intentie en wijze hoe een uitkomst tot stand gekomen is kenbaar gemaakt kan worden aan een belanghebbende.*
Hoe houden jullie rekening met verlies of verandering van de toestand van een onbemenste toepassing? (*'concept drift' of 'model drift'*) Licht toe.	*Over tijd kan het zijn dat een OT door updates of andere wijzigingen is veranderd ten opzichte van het tijdsstip van de totstandkoming van een uitkomst. Denk aan maatregelen zoals: versiebeheer, repository, alerts etc.*
Waar leggen jullie alle factoren om een uitkomst van een onbemenste toepassing te reproduceren vast? Is deze locatie altijd beschikbaar?	*Factoren van een OT zoals code, datasets en complementaire documentatie dienen ergens opgeslagen en vastgelegd te worden (zie aandachtspunt 1) om een uitkomst te kunnen reproduceren. Bijv.: wordt dit vastgelegd in een klantdossier, repository, in de cloud etc.*
Hoe reproduceren jullie een uitkomst bij een learning-based model?	*Een model op basis van leren maakt reproduceerbaarheid soms moeilijk tot onmogelijk. Welke maatregelen hebben jullie ingesteld om reproduceerbaarheid alsnog te faciliteren? Bijv. virtual machine, repository, versies van modellen etc.*

3.2.2 Aandachtspunt 4: Operational Recordkeeping / Monitoring

Herleidbaarheid is niet alleen van toepassing op de uitkomsten, maar ook op de werking van systemen. Herleidbare onbemenste toepassingen moeten hun gedrag op een systematische wijze registreren. Maar wat er geregistreerd moet worden, hoe en hoe lang zijn nog zeer context gebonden vragen. Het is belangrijk dat bij het ontwerpen van een onbemenste toepassing rekening wordt gehouden met het uiteindelijke toepassingsgebied als de technologie die gebruikt wordt. Het bijhouden van records, oftewel gestructureerde logs, zorgt voor een robuuste basis waarop een uitkomst/ resultaat herleid kan worden. Het implementeren van gestructureerde logs ("operational recordkeeping") of een monitoring systeem is dan ook één van de belangrijkere aandachtspunten.

In het voorstel voor de opkomende AI-act staat ook een clausule over "Record-keeping":

Artikel 12.4: Voor AI-systemen met een hoog risico als bedoeld in punt 1, a), van bijlage III voorzien de loggingcapaciteiten ten minste in
De registratie van de duur van elk gebruik van het systeem (begindatum en -tijd en einddatum en -tijd van elk gebruik);
De referentie database aan de hand waarvan de inputdata zijn gecontroleerd door het systeem
De inputdata ten aanzien waarvan de zoekopdracht een match heeft opgeleverd;
De identificatie van de natuurlijke personen die betrokken zijn bij de verificatie van de resultaten, zoals bedoeld in artikel 14, lid 5

In de Artifical Intelligence Act zijn in Artikel 12 clausules opgenomen voor record-keeping (registratie). Dit artikel stelt dat een AI-systeem ontworpen en ontwikkeld moet worden met de capaciteit van automatische registratie van gebeurtenissen (logs) tijdens de werking van het AI -systeem. Deze loggingcapiciteiten moet een mate van herleidbaarheid waarborgen die passend is voor het beoogde doel van het systeem. In andere woorden, wat er "gelogd" moet worden hangt af van welk doel en welke context een systeem heeft. Zo moet er in een (geautomatiseerd) acceptatieproces veel meer relevante informatie gelogd worden dan in een simpeler vrijblijvende rekentool voor bijv. een autoverzekering 's premie. Tot slot stelt dit artikel een minimum[11] aan wat de loggingcapiciteiten in moeten voorzien.

Aaandachtspunten voor monitoring / operational recordkeeping:

Aandachtspunten	**Toelichting**
Aandachtspunt 4: Monitoring / Operational Recordkeeping	
Hoe worden onbemenste toepassingen gemonitord? Licht toe met een voorbeeld.	*Niet alleen Stichting Toetsing Verzekeraars vereist om een document die toelicht hoe OT gemonitord worden. Het draagt ook toe aan de herleidbaarheid doordat met monitoring 'concept drift' of 'model drift' voorkomen kan worden.*
Hoe worden processen en de stappen die genomen zijn om tot een uitkomst te komen geregistreerd en gedocumenteerd? Licht toe.	*Herleidbaarheid vereist dat een uitkomst terug getraceerd kan worden naar de stappen in het proces (lineage) als de gebruikte input. Hoe dit gedaan wordt verschilt per soort OT, denk aan: logging, trace links, provenance modellen etc.*

[11] Let op ! Dit artikel gaat vooral over AI systemen met een hoog risico en het minimum geldt dan ook voor een hoog risico. Echter kan het geen kwaad om al mimimaal aan deze punten te voldoen. Zie voor meer informatie de structuur van de AI act.

Welke (loggings) gegevens in het proces van de totstandkoming van een uitkomst leggen jullie vast?	*Vanwege de opkomende AI act, is het van belang dat er aandacht geschonken wordt wat er gemonitord en eventueel geregistreerd wordt. O.a. de opkomende AI Act vereist de registratie van: de duur van het gebruik, de referentie database, inputdata die ten aanzien van de zoekopdracht een match heeft opgeleverd en de identificatie van een natuurlijk persoon die betrokken is bij het resultaat.*

3.3 Data

3.3.1 Aandachtspunt 5: Complementaire data documentatie

Een ander belangrijk aspect is het documenteren van de input data, met inbegrip van hoe deze data is verzameld alsook hoe de data is geselecteerd. Om dit op een gestandaardiseerde manier te doen, stellen wij het gebruik van *Datasheets for Datasets* (Timnit Gebru, 2018) voor.

In de elektra industrie wordt elk component gepaard met een datasheet die de karakteristieken, gebruik en andere informatie van dat component beschrijft. Geïnspireerd op dit idee stelt Gebru et al. (2021)[12] voor om bij elke dataset een datasheet aan te leveren. Een datasheet richt zich erop om de transparantie van de gebruikte data te vergroten door middel van een set aan vragen te stellen over de desbetreffende dataset. Deze set van vragen zijn verdeeld in verschillende categorieën, gebaseerd op een dataset lifecycle: motivatie, compositie, collectie, preprocessing/ cleaning / labeling, gebruik, distributie en onderhoud. Deze relevante vragen faciliteren uiteindelijk de documentatie van de input data, alsook hoe en waarom deze data is verzameld en geselecteerd.

Het standaard meeleveren van een gestandaardiseerd data document draagt ook bij aan het herleiden van datasets wanneer er sprake is van externe bronnen. Wanneer er bijv. van een externe partij een data set binnenkomt die gebruikt wordt in een onbemenste toepassing, kan met "Datasheets for Datasets" grotendeels herleid worden hoe deze data set bij de externe partij tot stand is gekomen.

Aaandachtspunten voor complementaire data documentatie.

Aandachtspunten	**Toelichting**
Aandachtspunt 5: Complementaire data documentatie	
Hoe documenteren jullie de wijze van de totstandkoming van de gebruikte (input) data?	*Herleidbaarheid vereist, expliciet bij learning-based modellen, geregistreerde documentatie over de wijze hoe/ waarom (input) data tot stand is gekomen en hoe/ waarom deze data is geselecteerd.*
Maken jullie gebruik van documentatie complementair aan elke dataset? Zo ja, hoe zit dit eruit?	*Het gebruik maken van gestandaardiseerde complementaire documentatie draagt bij aan het kunnen herleiden van de totstandkoming van een dataset. Denk aan: datasheets for datasets, data nutrition labels etc.*
Bij het gebruik van externe databronnen, of het leveren van databronnen aan andere partijen, hoe wordt de totstandkoming van de dataset uitgelegd? Welke documenten worden bij de dataset meegeleverd?	*Om goed te kunnen herleiden, is het zeker bij het gebruik van data afkomstig van externe bronnen, het belangrijk om te weten hoe en waarom de datasets tot stand zijn gekomen. Ook in het geval van uitgaande datasets. Er moet dan ook goed nagedacht worden op welke manier dit gedaan wordt. (Bijv. het meeleveren van een document gebaseerd op Datasheets for Datasets)*

[12] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://dl.acm.org/doi/10.1145/3458723

3.4 Uitkomsten

3.4.1 Aandachtspunt 6: Begrijpelijkheid

Het laatste punt dat wij graag onder de aandacht willen brengen is de begrijpelijkheid van de gedocumenteerde artefacten als de uitkomsten van een onbemenste toepassing. Herleidbaarheid heeft betrekking op het vermogen om de data, development en deployment processen van een geautomatiseerd systeem vast te leggen, gewoonlijk door middel van geregistreerde documentatie[13]. Enkel is het ook van belang dat deze geregistreerde documentatie begrijpelijk is voor (ondeskundige) belanghebbenden.

Mora Cantallops et al. (2021) noemt dit "Semantic Interoperability", oftewel de interoperabiliteit tussen de elementen en vereisten van herleidbaarheid en de bijbehorende context. In het geval van de verzekeringsindustrie, houdt dat dus in dat elementen van elke stap in het ontwikkelingsproces "gemapt" dienen te worden naar terminologie die relevant en begrijpelijk is voor de verzekeringsindustrie. Denk bijvoorbeeld aan de loggingcapiciteiten die niet alleen begrijpelijk zijn voor de programmeur, maar ook voor de (niet technisch geschoolde) adviseur.

Een eerder onderzoek, uitgevoerd door Olivier Koster gaat dieper in op de uitlegbaarheid en begrijpelijkheid van onbemenste toepassingen. Zijn paper is te vinden op: https://www.sivi.org/actueel/publicatie-a-checklist-for-explainable-ai-in-the-insurance-domain/, en bevat een checklist die gebruikt kan worden om de uitlegbaarheid van een uitkomst te toetsen.

Aaandachtspunten voor begrijpelijkheid.

Aandachtspunten	Toelichting
Aandachtspunt 6: Begrijpelijkheid	
Is alle geregistreerde documentatie begrijpelijk en uitlegbaar voor derde partijen? Zie meer in een eerder onderzoek naar uitlegbaarheid.	*Herleidbaarheid heeft betrekking op het vermogen om de data, development en deployment processen van een geautomatiseerd systeem vast te leggen, gewoonlijk door middel van geregistreerde documentatie. Enkel is het ook van belang dat deze geregistreerde documentatie begrijpelijk is voor (ondeskundige) belanghebbenden.*

[13] Definitie van herleidbaarheid (zie appendix A: begrippenlijst)

4. Referenties

AI-HLEG. (2019). *ETHICS GUIDELINESFOR TRUSTWORTHY AI.* Brussels: European Commission.

Brundage, M. a. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.* arXiv.

KPMG. (2020). *Ethisch Kader Toolkit.* Opgehaald van Verzedke: https://www.verzekeraars.nl/media/8082/toolkit-ethisch-kader_def.pdf

Kroll, J. A. (2021). *Outlining Traceability: A Principle for OperationalizingAccountability in Computing Systems.* Naval Postgraduate School, Monterey.

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M. A. (2021). Traceability for trustworthy ai: A review of models and tools. *Big Data and Cognitive Computing*, *5*(2), 20. https://doi.org/10.3390/bdcc5020020

Koster, O., Kosman, R., & Visser, J. (2021, September). A checklist for explainable AI in the insurance domain. In International Conference on the Quality of Information and Communications Technology (pp. 446-456). Springer, Cham.

Olorisade, B., Brereton, P., & Andras, P. (2017). *Reproducibility in Machine Learning-Based Studies:An Example of Text Mining.*

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44). https://dl.acm.org/doi/abs/10.1145/3351095.3372873

Timnit Gebru, J. M. (2018). *Datasheets for Datasets.* doi:https://doi.org/10.48550/arXiv.1803.09010

5. Appendix A: Begrippenlijst

Onbemenste toepassingen (OT)
Een onbemenste toepassing is een toepassing die resultaten/uitkomsten levert die door geheel of gedeeltelijk geautomatiseerde processen tot stand komen. "Onbemenst" betekent dan ook dat er geen sprake is van tussenkomst van de mens. Omdat er op het moment nog veel discussie gaande is wat wel of niet onder "Kunstmatige Intelligentie" valt, hanteren we met de term "onbemenste toepassingen" een bredere scope dan enkel "Kunstmatige Intelligentie". Zo vallen zowel een simpele rekentool als een machine learning algoritme beiden onder onbemenste toepassingen.

Herleidbaarheid
De herleidbaarheid van een onbemenste toepassing heeft betrekking op het vermogen om de data, development en deployment processen van een geautomatiseerd systeem vast te leggen, door middel van gedocumenteerde, geregistreerde identificatie. (AI-HLEG, 2019)

Reproduceerbaarheid
Reproduceerbaarheid heeft betrekking op de mogelijkheid om een uitkomst, afkomstig van een onbemenste toepassing, te kunnen laten reproduceren door een onafhankelijke partij met behulp van benodigde artefacten (code, data, omgeving etc.). Er zijn verschillende gradaties van reproduceerbaarheid. Zo is *herhaalbaarheid* de hoogste mate van reproduceerbaarheid, waarbij een uitkomst met dezelfde code, data en bijbehorende documenten door de originele partij gereproduceerd kan worden. *Repliceerbaarheid* is de mate waarop een uitkomst gereproduceerd kan worden door een onafhankelijke partij.

Rule-based systemen
Rule-based systemen of modellen, ook wel "Business Rule Engine (BRE)" modellen genoemd, volgen het concept "Als X gebeurt, dan Y". In andere woorden, een rule-based systeem omvat een set aan afleidingsregels of business rules die afhankelijk van de input een bepaalde uitkomst genereert.

Learning-based systemen
Learning-based systemen, zijn systemen die leren op basis van historische input, maar zich ook aanpassen aan nieuwe situaties zonder menselijke interventie.