Validation of the MAB-EX Framework through Requirements Analysis and a Case Study

Bachelor's Thesis of

Arno Leue

At the KIT Department of Informatics
KASTEL - Institute of Information Security and Dependability
MASE - Research Group of Modeling and Analysis in Mobility
Software Engineering

First examiner: Jun.-Prof. Dr. Maike Schwammberger
Second examiner: Prof. Dr. Raffaela Mirandola
First advisor: M.Sc. Akhila Bairy

1st February 2025 – 2nd June 2025

Karlsruhe Institute of Technology
Department of Informatics
P.O. Box 6980
76049 Karlsruhe
Germany

Validation of the MAB-EX Framework through Requirements Analysis and a Case Study (Bachelor's Thesis)

I declare that I have developed and written the enclosed thesis completely by myself. I have not used any other than the aids that I have mentioned. I have marked all parts of the thesis that I have included from referenced literature, either in their original wording or paraphrasing their contents. I have followed the by-laws to implement scientific integrity at KIT.

During the preparation of this work I used DeepL (https://deepl.com) in order to improve the readability in the language of the thesis. After using this tool, I reviewed and edited the content as needed and take full responsibility for the content of the thesis.

Karlsruhe, 1st June 2025

. . . . . .

(Arno Leue)

Abstract

Autonomous software systems are increasingly making decisions and taking action in many areas of everyday life. With this, we let them take on increasingly more responsibility. Trusting them with this should require their behaviour to be transparent, understandable and comprehensible. However, this is becoming more and more complicated, especially with machine-learning-based systems. In order to ensure the necessary transparency and understandability, these systems should be capable of explaining their behaviour themselves.

To this end, the Monitor, Analyze, Build, Explain (MAB-EX) framework has been developed and is continuously being improved as a reference framework. Since the framework is still defined fairly coarse, this thesis compares requirements for self-explainable systems from architectural- and user-perspectives. The goal is to analyze whether the current state of the framework is sufficient or if additional aspects should be incorporated.

From an architectural perspective, a requirements analysis based on a literature review identified 77 requirements focused on the idea of MAB-EX. While not all are strictly necessary for systems to achieve self-explainability, meeting all of them may result in higher-quality explanations. From a user-perspective, through conducting a case study of a Transport Management System (TMS), 21 more abstract requirements were derived by analysing needs indicated in user stories.

A comparison of both sets of requirements revealed aspects that could be of use for MAB-EX: Apart from some very technical requirements, all architectural requirements found could be confirmed by user-requirements. Since the similarities between the sets are bidirectional, all user-requirements were also covered by architectural requirements inversely. However, when considering only architectural requirements that were explicitly incorporated into MAB-EX to date, not every user-requirement was covered. Therefore, for MAB-EX to be considered a complete framework, non-covered requirements may need to be incorporated.

Zusammenfassung

In vielen Bereichen des alltäglichen Lebens entscheiden und handeln Softwaresysteme zunehmend autonom. Damit überlassen wir ihnen auch zunehmend mehr Verantwortung. Um ihnen dabei vertrauen zu können, müssen ihre Handlungen transparent, verständlich und nachvollziehbar sein. Dies wird jedoch immer komplizierter, speziell auch durch Systemkomponenten, die auf maschinellem Lernen basieren. Damit aber die nötige Transparenz und Verständlichkeit gewährleistet werden kann, müssen diese Systeme ihr Verhalten selbst erklären können.

Zu diesem Zweck wurde das Monitor, Analyze, Build, Explain (MAB-EX) Referenz-Framework entwickelt und kontinuierlich verbessert. Da das Framework jedoch noch relativ ungenau ist, vergleicht diese Arbeit Anforderungen an solche selbst-erklärbaren Systeme aus architektonischer und Nutzersicht. Das Ziel dabei ist herauszufinden, ob der aktuelle Stand des Frameworks ausreicht, oder ob zusätzliche Aspekte eingearbeitet werden sollten.

Durch eine literaturbasierte Anforderungsanalyse mit einem Fokus auf MAB-EX wurden auf architektonischer Seite 77 Anforderungen identifiziert. Zwar sind nicht alle davon notwendig, um Systemen Selbsterklärbarkeit zu ermöglichen, doch ihre Erfüllung würde zu qualitativeren Erklärungen führen. Auf Nutzerseite wurden durch eine Fallstudie eines Transport-Management-Systems 21 allgemeinere Anforderungen abgeleitet, indem Funktionalitätsbedarfe aus Nutzer-Stories extrahiert wurden.

Durch einen Vergleich beider Anforderungsmengen konnten Aspekte gefunden werden, die für MAB-EX von Nutzen sein könnten: Abgesehen von einigen sehr technischen Anforderungen konnten alle gefundenen architektonischen Anforderungen durch Nutzeranforderungen bestätigt werden. Da diese Ähnlichkeiten zwischen den Mengen bidirektional sind, wurden auch umgekehrt alle Nutzeranforderungen durch architektonische Anforderungen abgedeckt. Betrachtet man jedoch nur die architektonischen Anforderungen, die bisher explizit in MAB-EX eingearbeitet wurden, so decken diese nicht alle Nutzer-Anforderungen ab. Diese nicht abgedeckten Anforderungen sollten daher noch in MAB-EX integriert werden, wenn dieses den Anspruch hat, ein vollständiges Framework darzustellen.

Contents

List of Figures

List of Tables

List of Requirements

Theory-Requirements:

Case-Requirements:

List of Acronyms

Acr.	*Term*	*Page of Definition*
A	Addressee	3
AI	Artificial Intelligence	
C	Context	3
CM	Context Model	20
CPS	Cyber Physical System	3
CR	Case Requirement	
CS	Control System	16
E	Explanation	3
EM	Explanation Model	7
ES	Explain System	16
G	Target Group	3
H	Stakeholder	3
I	Information	
IEEE	Institute of Electrical and Electronics Engineers	
LLM	Large Language Models	
M	Means	5
MAB-EX	Monitor, Analyze, Build, Explain Framework	6
MAPE	Monitor, Analyze, Plan, Execute Feedback Loop	
ML	Machine Learning	
N	Need	3
NFR	Non Functional Requirement	15
O	Explainability Goal	4
Q	Explanation Quality	10
R	Representative	3
RE	Requirements Engineering	
RQ	Research Question	
S	System	3
SE	Software Engineering	
SLR	Systematic Literature Research	
SM	System Model	19
SXA	Self Explainability	5

SXS	Self Explainable System
TMS	Traffic Management System 29
TR	Theory Requirement
UE	Usability Engineering
X	Explanandum 3
XA	Explainability 4
XAI	Explainable Artificial Intelligence
XE	Explainability Engineering 5
XR	Explainability Research
XS	Explainable System 4
XT	Explainability Theory 9
Y	Aspect 3

1. Introduction

We live in the age of Artificial Intelligence (AI). With the recent gain in popularity of Large Language Models (LLM), software systems based on Machine Learning (ML) have reincreased the attention on (intelligent) Cyber-Physical System (CPS) in the past years. At the same time, their decisions have become more powerful, the decision-making processes more complex and thus the systems more opaque for all participants involved [7]. "Due to this complexity, it becomes increasingly difficult for system- and software-engineers, but also users, auditors, and other stakeholders, to comprehend the behaviour of a system." [9] Consequently, it will become essential for Cyber-Physical Systems to explain their behaviour to ensure user trust and understanding of them [9, 10].

The ability to explain the behaviour or decisions of such a system, i.e. explainability, sometimes referred to as interpretability, understandability or transparency [10, 11], may be implemented into the system of interest, possibly resulting in a self-explainable system (SXS). This ability can be seen as a requirement imposed on the system by the stakeholder involved with it. Such requirements describe representations of perceived needs of stakeholders or properties or capabilities that a system should have, and thus form the basis of good quality software. In other words: Software quality is the degree to which a system fulfils the stated and implied requirements of the needs of the stakeholders [8]. Explainability can be considered one of those quality requirements.

The branch of Software Engineering (SE) that addresses requirements is called Requirements Engineering (RE). Similar to Usability Engineering (UE) describing the branch of RE that focuses on improving the usability of interactive systems, Explainability Engineering (XE) focuses on improving the explainability of a system [12]. The most prominent branch of XE is Explainable Artificial Intelligence (XAI), which focuses on "expert explanations for very complex AI algorithms" [1]. Explainability Engineering, on the other hand, seeks to provide holistic explanations for entire systems and a variety of stakeholders by combining and preparing information extracted from different system-components [1].

Although the XE research area is still quite new, a first modular reference framework for the integration of self-explainability into software systems has been defined [9] and extended [1]. Its name, Monitor, Analyze, Build, Explain Framework (MAB-EX), and its roots are based on the Monitor Analyze Plan Execute (MAPE) feedback loop defined by Kephart et al. [13], which is referred to as "the most influential reference control model for autonomic and self-adaptive systems" [14]. While some of the research questions posed the authors of MAB-EX have been addressed and preliminary answered [12], the framework is still very coarsely defined [1] and thus impractical for implementation.

In order to bridge this gap in detail and to validate whether the current state of MAB-EX is sufficient as a reference framework for building self-explainable systems, this work compares requirements imposed onto a system that might implement the framework, from a theoretical-general and a practical-specific point of view.

For this purpose, the results of a theoretical literature review are summarized in Section 2. This research is orientated towards an understanding of the state of the art in the field of Explainability Engineering as opposed to the specifics of XAI, in order to keep the perspective of all possible systems implementing MAB-EX. The results are then used in Section 4 to phrase research questions. In a theoretical requirements analysis of the literature in Section 5, architectural requirements onto self-explainable systems are elicited and explained. In a subsequent practical case-study in Section 6, stakeholder needs are elicited as user-requirements. A mutual coverage of both sets of requirements is then analysed in Section 7 in order to answer the research questions.

2. Background

A system (S) is a collection of components organized to accomplish specific functions [15]. Systems can be classified according to various aspects. One such aspect is the type of system-components, e.g., software or physical, resulting in the class of Cyber-Physical-Systems (CPS), a combination of both. In more detail, a CPS is a collection of independently interacting components, that primarily transmutes how we interact with the physical world. [16] When referring to *systems* in this thesis, CPS are meant.

The following provides an overview of fundamentals that are helpful for this thesis. First we will start with basics of *explanations* in Section 2.1 and *explainability* in Section 2.2. Then, we will cover the research field of *Explainability Engineering* and artefacts of it in Section 2.3, especially the MAB-EX Framework in Section 2.4.

2.1. Explanations

An explanandum (X), also known as a *phenomenon* [17], is an entity to be explained. This can be a system in general, or a specific aspect (Y) of it, such as its reasoning processes, inner logic, model's internals, intention, behaviour, decision or knowledge [6].

The need (N) for an explanation of that explanandum originates in a lack of *understanding* of it [4] by some entity. Thus, an addressee (A), also known as stakeholder (H), recipient or explainee, is required, i.e. a person or domain of expression [4, 9].

Given that such a need is often expressed more than once in a similar manner, these types are often classified into target groups (G). Explanations that address this common need are then being formed so that they fit to a representative (R) equipped with the background knowledge and processing capabilities characteristic of the group [4]. *Explainers* refer to a system or specific parts of a system that supply its stakeholders with the needed information [6].

A context (C) of an explanation is set by a situation consisting of the interaction between an addressee, a system, a task, and an environment [6].

An explainability goal (O) of an explanation for a stakeholder group is a target to be reached by the process of explaining to the stakeholder.

The literature examined contained different definitions of explanations (E) and explainability. Although a detailed comparison of these definitions would be relevant, it is beyond the scope of this thesis. Condensed to their core concepts, definitions of explanations could be compared as seen in Table 1 with the key differences in their intended purpose.

Table 1: Comparison of different definitions of Explanations.

Definition An explanation E for X is a piece of information I that ...	Authors	Year
... makes R of G understand X with respect to Y in C.	Köhl et al.	2019
... makes X interpretable by G.	Bersani et al.	2023
... contributes to A's understanding of X in C.	Chazette	2023
... makes X understandable by G with respect to O.	Schwammberger et al.	2024

According to these definitions[1], the subsequent working definition is proposed for the purpose of consistency within the scope of this thesis. It is derived on the definitions of Köhl et al. [4], Schwammberger et al. [1] and Chazette [10] combining the element of explanations as contributions with the taxonomy of relevant elements such as the explainability goal and aspects.

Explanation **Definition 1**

Information (I) that contributes to a representative (R) of a target group (G) understanding an aspect (Y) of an explanandum (X) in a context (C) with an explainability goal (O).

2.2. Explainability

As already mentioned, Köhl et al. stated that "explainability is not a technical concept but tightly coupled to human understanding" [4] and that "what makes a system explainable is the access to explanations" [6]. As that, understanding must be included in the definition of explainability of systems as well.

Explainable Systems

To add the ability to explain explananda, i.e. explain-ability (XA), to systems, all authors of the mentioned definitions added means to an end of being able to produce explanation. Condensed to their core concepts, definitions of explainability were phrased as seen in Table 2, with smaller differences in their wording.

Table 2: Comparison of different definitions of Explainable Systems.

Definition S is explainable if and only if ...	Authors	Year
... M is able to produce an E of X regarding Y, for G in C.	Köhl et al.	2019
... T enables A to understand X of S by giving I in C.	Chazette et al.	2021
... M is able to produce an E of X for G in C.	Bersani et al.	2023
... M is able to produce an E of X with an O for G in C.	Schwammberger et al.	2024

[1]Full definitions in Proposition 2 – Proposition 5.

Based on these, a second working definition is proposed below, again combining taxonomies of relevant aspects and adopting the more common descriptor means (M) for the instance providing the explaining information:

Explainability Definition 2

Ability of means (M) to produce an explanation (E) of an aspect (Y) of an explanandum (X) of a system (S) in a context (C) for a representative (R) of a target group (G).

In contrast to some previous definitions of explainability [6, 10], the working definition does not actively state the target group's goal of understanding the explanation, since the reference to explanations already includes this in a transitory way.

Self-Explainability

In order to achieve the goal of self-explainable Cyber-Physical Systems [9], this thesis also follows common definitions of self-explainability by requiring the previously defined means, that produce an explanation, to be within the system that showed the explananda.

Self-Explainability Definition 3

Ability of means (M) of a system (S) to produce an explanation (E) of an aspect (Y) of an explanandum (X) of the same system (S) in a context (C) for a representative (R) of a target group (G).

This is in line with the definitions by the aforementioned authors, matching the syntax by Bersani et al. [7] and Schwammberger et al. [1], and rephrasing the content of Buiten et al. [11] stating that "a self-explainable software system is one that provides explanations for its (own) (...) behaviour".

2.3. Explainability Engineering

As already motivated, the branch of Software Engineering (SE) that addresses requirements based on the explainability of systems is called Requirements Engineering (RE) [10], and a branch of that, Explainability Engineering (XE), focuses on improving the explainability of a system, therefore including SE and RE issues [12].[2] The most prominent branch of XE is Explainable Artificial Intelligence (XAI), which often only focuses on explanations for AI algorithms, whereas the goal of XE is to combine information from different systems and to deliver it as explanations to different stakeholders [1].

[2] See also: Figure 4.

Next to XAI, other prominent research topics of Explainability Engineering have already been phrased by Brunotte et al. [12], such as *technical aspects* of explanation generation, the *design process* and *quality aspects* of explainable systems, and *resources* of explainability including possible *costs* and useful *artefacts*.

Artefacts in Explainability Engineering

Artefacts are "textual or graphical documents with the exception of source code" [18] varying in size and granularity that are typically (re-)used as guidance during Software Engineering activities, or to gather knowledge in general [10]. Important artefacts of discussion for this thesis are:

Reference Models "A reference model is an abstract framework for understanding significant relationships among the entities of some environment. It enables the development of specific reference or concrete architectures using consistent standards or specifications supporting that environment. A reference model consists of a minimal set of unifying concepts, axioms and relationships within a particular problem domain, and is independent of specific standards, technologies, implementations, or other concrete details." [19]

Frameworks "A framework generally provides a skeletal abstraction of a solution to a number of problems that have some similarities. A framework will generally outline the steps or phases that must be followed in implementing a solution without getting into the details of what activities are done in each phase." [20]

2.4. The MAB-EX Framework

The Monitor, Analyze, Build, Explain Framework is the "first reference framework for building self-explainable systems that can provide explanations about the system's past, present, and future behaviour" [12]. For this purpose, an additional (sub)system analyses the behaviour of the system showing phenomena of interest. Thus, a distinction is often made between the explain- and the control system parts. Hence its name, the framework is built on four major phases in the creation of explanations at runtime:

1. Phase: **Monitoring** the software system, its environment and possibly the addressee. For this, relevant sensor data, commands from the system, stakeholder-system interactions and former explanations are captured in chronological order, possibly as logs. [9]
2. Phase: **Analysing** the monitored data to detect an explanation need, which can either be triggered externally by an addressee or internally by the system. The latter happens, when the system detects deviations from observed behaviour that might indicate an explanation need. [9]

3. Phase: **Building** an explanation by evaluating an internal explanation model (EM) of the system, that captures causal relationships between events and system reactions. It allows for identifying possible causes for the behaviour that needs to be explained. It may also allow for look-ahead simulation. [9]
4. Phase: **Explaining** the phenomena to the addressee by transferring the built to an understandable explanation for the target group. Therefore, it should be target-specific. To this end, we use a mental model with a human addressee or an interface otherwise. The model describes preferences of the recipient w.r.t. possible explanations. [9]

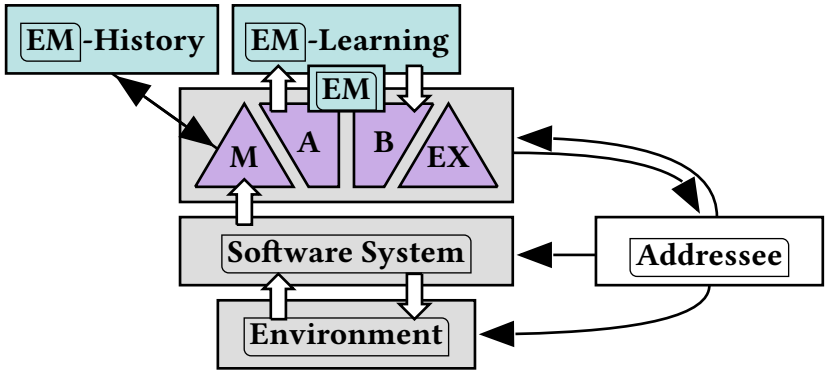

Figure 1: The MAB-EX Framework in its current version by Schwammberger et al. [1], visually adapted. The four phases are highlighted in purple, entities connected to an explanation model used in the phases in teal.

Unlike the MAPE-loop, on which the framework is based, loops may not occur because another monitor phase is not necessarily triggered again. Though, a loop can occur by monitoring the addressee for its reaction or by explicitly needing a follow-up [9].

3. Related Work

Besides the release of MAB-EX, a lot has happened in Explainability Engineering over the last few years. In order to not only derive requirements from the framework, but also to include more aspects of explainability, we will summarize some findings before going into the main research. For the sake of simplicity, the term Explainability Theory (XT) will be used to refer to the research of explanations and explainability in general.

Definitions in XT As already discussed in Section 2.1 and Section 2.2, the research field of defining explainability has brought forward different definitions [1, 4, 6, 7, 10]. In more detail, as Brunotte et al. [12] already examined, Köhl et al. [4] and Chazette et al. [21] explored explainability from the perspective of Requirements Engineering, based on which Chazette et al. [6] proposed a more detailed definition.

MAB-EX Publication In its publication the authors of the framework name a series of challenges in the vision of "truly comprehensible, flexibly tailored explanations" [9]. These include, among others: exterior challenges such as the *comprehensibility*, *presentation* and *interactivity* of explanations, and interior ones such as explaining post-hoc what happened or in advance, what might happen - based on predictions.

Explanation Need Cases Sadeghi et al. [5] proposed categories and a taxonomy of situations, they refer to as cases, that need explanations. Additionally, they give examples for these types of cases in order to be used as "to guide the requirements elicitation for explanation capabilities of (…) systems". The authors see the proposed categories "as a means to extract (…) explanation situations" at design time of a system.

Explanation Model Building Schwammberger et al. [2] worked on the idea of explanation models, and proposed an extraction and refinement process for generating them out of system and context models.[3] They proposed that the resulting process could be incorporated into the build phase of MAB-EX.

Explanation Timing Bairy et al. [22] addressed the timing of explanations as a crucial aspect. They envision to find the optimal timing of an explanation either before, during or after an explanandum based on the attention level of the recipient.

[3] More Details are in the Appendix.

Explanation Requirements Chazette et al. [23] proposed a "quality framework that links the dependencies, characteristics, and evaluation methods for explainability-requirements". Their results of a case-study showed that the framework is "beneficial in an industrial context and leads to the construction of explanations that increase usage frequency, system acceptance and user satisfaction" [23].

Levels of Explainability To classify a system's explainability, Bersani et al. [7] proposed four levels.[4] The ability to just recognize a need for an explanation at level 2, the additional ability to produce a local or global explanation at levels 3 and 4 and none of the above at level 1. A fifth and highest level, in which the generation of explanations is performed decentrally in communicating feedback loops, has been proposed [24], but subsequently been included as additional dimension in levels 3–4.

Explanation Satisfaction In order to determine whether an explanation has satisfied its addressee, Bersani et al. [7] also proposed a measure of explanation quality. They formalize an explanation as an object dependent on its influencing factors, in their case by its level, explanandum, context and target group: $E(L_i, X, C, G)$. With this, they define an orderable metric dependent on a means: $Q_E(M)$. In order for the explanation to be satisfying, this metric has to be greater than a threshold ε proposed by the corresponding target group. They call this the *Explainability Requirement* R_E.[4]

Explanation Benefit Schwammberger [3] proposed a need and definition for explanation correctness and goodness.[4] They argue that "trustworthiness of explanations can only be reached through a holistic Explainability Engineering process", which additionally complicates the achievement of this need.

MAB-EX Adaption Schwammberger et al. [1] integrated the before-mentioned explainability-levels by Bersani et al. [7] into MAB-EX. They propose to investigate additional levels, e.g. "whether an explanation is time critical or not". Additionally, they took onto the definition of explanation quality proposed by Bersani et al. [7].[4]

Apart from the explanation-timing, Bairy et al. [22] also gained insights on explanation-justification. Winikoff et al. [25] proposed a framework for constructing explanations and presented an algorithm for giving these. Buiten et al. [11] examined explainability and changes to requirements of autonomous systems in legal environments.

[4]More Details are in the Appendix.

4. Research Design

In order to address the question of the practicality of the MAB-EX Framework, it is first necessary to look at what the overall goal of self-explainable systems, with a focus on MAB-EX, yields in terms of research needs. Therefore, the following defines these. After that, research questions are stated on the basis of these definitions.

Research Needs

A general research need in the area of self-explainability is the need for this ability in systems themselves. It arises from the desire to be able to rely on these systems or to be able to trust them.

> ***Research Need 1:*** *The need for self-explainable systems.*

From a developer's point of view in Software Engineering, general functionality such as a system's ability to generate explanations should follow some standards in practice. Thus, the need for self-explainability implies a need for some guidelines as a standard for building such systems:

> ***Research Need 2:*** *The need for a guideline for building self-explainable systems.*

As a framework, MAB-EX consists of defined phases that should be followed in order to implement a solution for self-explainability of Cyber-Physical Systems. However, in order to implement this reference framework in software systems, one could phrase architectural implications of it as general requirements. These can be used as a guideline in an implementation phase in Software Engineering, as a "set of all requirements forms the basis for subsequent development of the system (...)" [15]. In this case, these requirements should not include details on how a system looks like in practice, in order to fit to every possible implementing system. For the development, we used the following definition of requirements, suggested by the IEEE standard glossary:

Proposition 1 *(Requirements by [15]): (1) A property or capability required by a stakeholder to solve a problem or achieve an objective. (2) A property or capability that must be met or possessed by a system or system-component to satisfy a contract, standard, specification, or other formally imposed document. (...)*

On this basis, the following third research need can be expressed, focusing on the previously explained need for a set of requirements for a MAB-EX implementation:

> ***Research Need 3:*** *The need for a set of requirements for implementing MAB-EX.*

Research Questions

Requirements can be divided into subgroups because they can describe different needs for systems. As described in Proposition 1, this can be user-requirements or requirements based on the technical architecture of a system [15]. To this end, one can try to validate already proposed requirements by evaluating user needs. In order to answer the question of whether MAB-EX is already suitable for implementations or still too coarsely defined, one could phrase the following research question:

> ***Research Question 1:*** *Can (the state of) MAB-EX be validated by comparing Explainability-Requirements from theory and practice, i.e. architectural and user side?*

With the artefact of a list of all evaluated requirements as a guideline, one can also try to answer a second research question based on Research Need 2 and Research Need 3:

> ***Research Question 2:*** *What Requirements are needed for the implementation of self-explainable systems (at run-time, overall, etc.) using MAB-EX?*

Research Approach

To answer these questions, theoretical requirements for explanations based on MAB-EX, but also on Explainability Theory, social sciences and philosophy, are stated. Many of them are based on the elicitation of fundamentals found within a literature research in Section 2 and Section 3. Research Question 2 is answered at the end of Section 5.1, where the requirements elicited are also visualized.

Subsequently, the requirements based on user needs should be examined in a case study in Section 6 using a concrete example scenario. The requirements should be specific to the scenario and thus describe practical requirements for the example system. By specifying explanation-cases, different contexts for explanations are described, from which the requirements are then derived.

Afterwards, both requirements should be checked for coverage within an evaluation in Section 7. Here the focus should be on Research Question 1.

Artefacts

With the above structure, this thesis proposes the following types of artefacts: definitions in explainability, a knowledge catalogue of requirements, and explanation-cases for self-explainable systems, all but the latter with a specific focus on MAB-EX.

Definitions are of importance for a given topic, concept, or aspect. They tend to evolve and change over time, as new insights emerge, and vary according to the focus of the field of study. They should be agreed upon between individuals in a given context. [10]

Knowledge Catalogues document knowledge about a given topic. They are usually used during Requirements Engineering. [6]

Explanation-Cases are scenarios that describe a concrete situation in which a software system should provide an explanation. They illustrate demands for explanations. [5]

To the best of our knowledge, there has been no contribution that specifically addresses this type of research for MAB-EX, or for any other self-explainable system framework.

5. Requirements Analysis

As a general overview, the requirements elicited in the following will be classified by different aspects. As in Requirements Engineering, requirements are generally classified according to two criteria: their necessity to be implemented (optional/mandatory) and whether they describe specific functions of a system-component or broader constraints onto the whole system (functional/non-functional).

Optional (O) requirements, as their name suggests, are not mandatory.

Functional (F) requirements, also known as capabilities [10], "specify functions that a system or system-component must be capable of performing" [15]. Examples include "formatting text, calculating a number, modulating a signal" [10].

Non-Functional (NF) attributes of or constraints on a system. They are usually detailed statements of the conditions under which the solution must remain effective, qualities that the solution must have, or constraints within which it must operate. [10]

Additionally, requirements are marked with the phase of MAB-EX in which they are located and needed (M, A, B or X). However, not all requirements can be tied to a specific phase. This includes premises (P), which require the existence of entities or objects needed to explain something; details on the explanation model used in MAB-EX to build explanations (E); and general (G) requirements that cannot be tied to a specific phase:

Premises (P) based on the explanation environment.

Phase (M, A, B, X) requirements specifically linked to the four phases of MAB-EX: monitoring (M), analysing (A), building (B) and explaining (X).

Explanation Model (E) requirements related to the design of the explanation model that is built and used in MAB-EX.

General (G) requirements on explainable-systems, that cannot be tied to another above.

As an aid, all requirements that have been explicitly mentioned in papers defining or improving MAB-EX are additionally marked with (MAB-EX) .

5.1. Elicitation: research analysis

In line with MAB-EX's goal of creating self-explainable systems [9], this elicitation will start with the basic assumptions defined by explainability. Dependencies between requirements will be explained as comprehensively as possible, but due to their complexity, this is not always possible. An overview of all connections can be found in Figure 2.

5.1.1. Base Environment Assumptions

For the following analysis, we will combine Definition 1 of explanations and Definition 2 of explainable systems. Additionally we will require the system to be self-explainable (Definition 3), i.e. the means explanation-production have to part of the system. This combination translates to at least six clauses that must be present for self-explainable systems to fulfil this goal.

Definition 1, 2 and 3, combined and fragmented into clauses:
(a) Ability of [means (M)] → [P2]
of a [system (S)] → [P1]
(b) to produce [information (I)], → [X9]
(c) with respect to an [explainability goal (O)] → [P3]
(d) that contributes to a [representative (R)] of a [target group (G)], → [P6]
(e) understanding an [aspect (Y)] of an [explanandum (X)] → [P5]
of the same [system (S)]
(f) in [context (C)]. → [P4]

System

Combining (a) and (b) with Definition 3, results in two requirements: First, we need to have a system (S), and second, it has to have a (sub)system M that needs to be able to produce an explanation. Thus, for the rest of this thesis, we will refer to M, as a part of S, as the [explain system (ES)] and the remainder of S without M to as [control system (CS)].

Presence of Software System: **P1**
A system with software components must exists.
Presence of Means: **P2**
The system must have means to produce an explanation [1, 4, 7, 10].

Additionally, as stated in (c), with the production of an explanation, the explain system follows some explainability-motivation or -goal (O), that it strives to achieve. (In many cases, this can be summarized as trust in the control system, its behaviour or decisions.)

MAB-EX **Presence of Explainability Goal:** **P3**
The explain system's content must have an explainability goal. [1]

Context

The remaining clauses (d)-(f) deal with the context (C), in which an explanation is given. This context can be divided into entities, the most important of which are stakeholder (H) (or addressees, representatives) and the explananda (X) of interest to H. As requirements we express the existence of C in general and the existence of H and X in particular:

MAB-EX **Presence of Context:** **P4**
The control system has some context in general. [1, 4, 7, 10]

MAB-EX **Presence of Phenomena:** **P5**
The control system shows phenomena in its context. [1, 3, 4, 7, 10]

Presence of Stakeholders: **P6**
The control system has different stakeholders.

Additionally, we can refine the stakeholder-presence more detailed: Overall, one can state that H has some kind of goal when using the explain system. Thus, when a phenomenon happens, a need can arise in H, which can lead to the production of an explanation.

Presence of Stakeholder Goal: **OP7**
A stakeholder should have a goal when using the explain system. [26]

Presence of Explanation Needs: **P8**
The control system's phenomena cause different explanation needs.

5.1.2. The Aspect of Understanding

On the basis of Definition 1, the 'presence of means to produce an explanation' P2 should be sub-divided into two additional requirements: The ability to just produce information, and the aspect of contributing to understanding with it, which makes the information an explanation.

MAB-EX **Producibility of Content:** **X9**
The explain system must be able to produce information. [4, 9]

Benefit of Content: The information produced should contribute to a stakeholder understanding explananda of the control system. [3, 27] **G10**

The former results in a requirement that summarizes and motivates the explaining phase (EX) of MAB-EX, the latter motivates the use of self-explainable systems.

Contributions to Understanding

To pick up on the point of contribution, one aspect that is necessary for information to contribute to a stakeholder's understanding of the explanandum is the understanding of the information provided itself.

Understandability of Content: **G11**
The information produced must be understandable by a stakeholder.

Because in order for a stakeholder (H) to make use of information, H must first understand the meaning of it, and secondly incorporate it into its world model. The latter won't be covered in detail for now, but the former by dividing it into the pure understanding of the information-structure and the understanding of its meaning:

Comprehensibility of Content: The information's presentation should be understandable by a stakeholder. [27] **G12**

Grasping of Content: The information's meaning should be understandable by a stakeholder. **G13**

Adaptability to Stakeholder

Thus, in order to present the information in a way a stakeholder can understand, it must be fundamentally adaptable by the explain system (ES), and the ES must be able to know which information is appropriate for that stakeholder:

Adaptability of Information: **G14**
The information should be adaptable during production.

Preferences of Stakeholder: The explain system should be able to build appropriate displays of information for stakeholders. [23, 26, 27] **B15**

Based on G14, requirements can now be stated for the *EXplaining* phase of MAB-EX, which is responsible for producing information based on the explain system's knowledge.

Adaptability of Information: Syntax and Semantics

Using classifications of explanation-details by Chazette et al. [23], Nunes et al. [26], Nauta et al. [28] and Kahn et al. [27], information intended as an explanation can be adapted in many ways, most commonly related to its presentation. Consequently, in order to combine the differentiation of understanding into representation G12 and meaning G13 of information with its adaptability G14, the following requirements call for an external and internal adaptability:

MAB-EX

Adaptability of Presentation Syntax: The syntax of information presentation must be adaptable. [9, 10, 23, 26–28] **X16**

Adaptability of Presentation Semantics: The semantics of information presentation must be adaptable. [23, 26–28] **X17**

The surface presentation, i.e. the syntax of the information, can be adapted in a number of ways, some important examples of which are the media type (e.g. textual, visual, audio), the amount of media used in parallel (e.g. videos: visual + audio) and the size of it (e.g. the length of a video in minutes):

Adaptability of Media: The medium, in which information is expressed, should be adaptable. [23, 26] **OX18**

Adaptability in Media-Amount: **OX19**
The amount of different media types should be adaptable. [23]

Adaptability in Amount: **OX20**
The *size* (or length) of information should be adaptable. [27, 28]

Similarly, the semantics of the information-presentation can be adapted in many aspects, including the relevance of the information provided regarding the phenomena of interest, its contrast, its complexity and the style of communication (e.g. factual, emotional, simplistic, paced). The contrast refers to the inclusion of information that is not directly aimed at, for example, *why A happened*, but also *why B did not*:

Adaptability of Relevance: The information should be of relevance to the phenomena of interest. [27, 28] **OX21**

Adaptability of Contrast: The information should be adaptable in its contrast, i.e. the scope of information included in it. [23, 28] **OX22**

Adaptability of Complexity: **OX23**
The information should be adaptable in its complexity. [27, 28]

Adaptability of Style: The communication of information should be adaptable in its style. [23, 26] **OX24**

5.1.3. Build of Information: Explanation Models

To continue with the second part of stakeholder-adaptability, that is, building information based on the preferences of a stakeholder (H), a model is needed that should capture both how and what information is presented in order for H to understand it and transitively get the best benefit from it, which would then fulfil the goals of the explain system and H.

MAB-EX **Buildability of Explanation-Model:** The explain system must be able to build a model as a basis for information production. [1, 2, 9] **B25**

This explanation model (EM) should represent the connections between context events and system-explananda in order to serve the explainability of the control system.

Explanation Model Build

Using the constraints on explanation models (EM) in the publication paper of MAB-EX by Blumreiter et al. [9] and in the more recent details on EM-generation by Schwammberger et al. [2], details on these are phrased in the following. The first is the ability to initialize it by extracting details from a model of the system. This system model (SM) can be an artefact from Software Engineering processes, e.g. architecture diagrams, communication protocols, etc. [3]

EM Initializability: **B26**
An explanation model should be extractable from a system model. [2]

Constraints on the use of a system model are that, as a basis for the information produced, it must contain every possible system behaviour correctly and accurately:

MAB-EX **Manual Constructability:** **OE27**
The explanation model should be manually constructable. [2, 9]

MAB-EX **Automated Constructability:** **OE28**
The explanation model should be automatically constructable. [2, 9]

Thereby the mapping from system models to explanation models should be possible both manually and automatically:

Complete Constructability: The system model should include every possible system behaviour. [2] **OE29**

Correct Constructability: The system model should include system behaviour accurately and correctly. [2] **OE30**

After initialization, the explanation model should be adaptable in its inclusion of information, i.e., irrelevant details should be hideable. This can go in two directions, making the explanation model more general or more individual.

EM Relevance: **E31**
Details not relevant for the explainability goal should be hideable. [2]

EM Generalizability: The explanation model should be generalizable towards a general target group. [2] **E32**

EM Individuality: The explanation model should be individualizable towards a specific representative of a target group. [2] **E33**

With Definition 1, and also in general, it makes sense to use both abilities, first to adapt the explanation model to general target groups, which then can be used as a basis to add specifics for representatives of these. Between generalization and individualization, the explanation model should be extensible with stakeholder-specific information from a context model to make the produced information 'context-aware'. This context model (CM), also called environment model *Env*, describes the operating context of a system. [3]

EM Contextualization: The explanation model should be extensible by stakeholder-specific information from a context model. [2] **E34**

5.1.4. Explanation Model Considerations

Connecting the requirements from the analysis of the understanding with the idea of an explanation model (EM), results in the conclusion that information extracted from the EM is generally needed to be linked to a stakeholder's knowledge and preferences.

Information Coherence: The information produced should be based on the stakeholder's knowledge base. [28] **G35**

But this adaptation also results in challenges, some important of which are not to include information that (a) tends to be biased or prejudicial to a stakeholder, and (b) it should not be privileged to receive.

Information Objectivity: The information produced shall be unbiased, unprejudiced and impartial. [27] **G36**

Information Security: The explanation model should be adaptable based on a stakeholder's privileges to ensure data security. [27] **E37**

The basis of the explanation model, the context model, has already been required to be correct and complete. Likewise, the information generated by the explanation model should also fulfil these two aspects. However, it should be specified that the generated information needs not to be complete, i.e. contain all explananda, but none relevant should be left out. Otherwise, a correct explanation could be achieved by removing all aspects that are not provably correct, which could result in it not being useful any more.

Information Correctness: Information generated shall be deduced from provably correct system models and context models. [3, 27, 28] **G38**

Information Completeness: Information generated needs sufficient breadth, depth and amount to fulfil explainability goals. [27, 28] **G39**

Combining these requirements results in a definition of information-goodness similar to that proposed by Schwammberger [3] in their research. However, this requirement cannot be met by an explanation model alone.

Information Goodness: Information must be correct and measurably help a target group to understand an explanandum. [3] **G40**

From the correctness of information, which is given by its deducibility from system- and context-models, it follows that in the case of the same system in the same context, i.e. with the same models, the same information must be deduced again. This can be described as consistency, or determinism. Therefore, it makes sense to require that similar information should be generated by similar models, i.e., in only slightly different conditions, explanation should not vary too much. This also increases the benefit for the stakeholder.

Information Consistency: The generation of information must be the same under the same conditions [27, 28], i.e. be deterministic. **G41**

Information Continuity: The generation of information should be similar, if the conditions change only slightly [28], i.e. be continuous. **OG42**

With the existence of an explanation model, it can now be assumed that all the information extracted from it is correct, but as touched on above, it is not yet certain that it contributes to an understanding of the system. Before addressing the latter, however, there are a few more aspects to consider regarding the usage of the explanation model.

5.1.5. Explanation Model Usage

Based on the requirement that information must be producible X9, a production based on the explanation model must be triggerable. It is advantageous if this can be done in two ways: First, by an explicit need expressed by a stakeholder in some way, and second, by an analysis of the context, including the stakeholder, that infers a need implicitly.[5]

MAB-EX **External Triggerability:** The production of information by the explain system must be externally triggerable (e.g. by stakeholders) [9]. **B43**

MAB-EX **Internal Triggerability:** The production of information by the explain system must be internally triggerable (e.g. by an analysis) [9]. **B44**

Therefore, first, a need must be generally detectable, and second, information must be extractable from the explanation model at all:

MAB-EX **Detectability of Explanation Needs:** **A45**
The explain system should be able to detect explanation needs. [7, 9]

MAB-EX **Extractability from Explanation Model:** **B46**
Information in the explanation model must be extractable. [2, 9]

Simulations

In order to be able to identify any need for explanations in advance and thus resolve them as quickly as possible, the explain system should be able to predict future explananda of the control system and their consequences for the understanding of stakeholders on the basis of the explanation model.

MAB-EX **Simulation of Behaviour:** Future explananda of the control system should be simulatable based on the explanation model. [9, 22] **B47**

It is therefore helpful to be able to use a history of past explanation models (EM) in connection to their contexts and to demand that the EM, and thus the basis of information, should always be kept as recent as possible. This means that it must be possible to update it at runtime and that it should always be up-to-date.

MAB-EX **History of Explanation Model:** The explain system should keep a chronological history of past explanation models. [1] **E48**

Information Recency: The information produced by the explanation model should be as recent (up-to-date) as possible. [27] **B49**

Updatability of Explanation Model: The explanation model adapted for a specific stakeholder should be updatable at run-time. [2] **B50**

[5] Note that it could be argued that the explicit expression of an explanation need is also just an irregularity of observed data. For example a voiced need is also just an 'irregularity' in the recorded language-data. In this case, only an 'Internal Triggerability' is needed.

In addition to recency, the timing of information should be considered in additional aspects as described in the following.

5.1.6. Timing of Information Production

In order to maximize the benefit of information with respect to the stakeholder's goal of understanding explananda (X), it should be possible to provide it before, during, or after X.

> **Information Timing:** An information should be adaptable in its timing, i.e. the occurrence of it before, during, or after an event [22]. **G51**

This then leads to the problem that it may be necessary to calculate benefits not only specifically for individual stakeholders, but also for several together. One way to tackle this is to require a parallel production and a hierarchy of priorities in production based on the importance of the information.

> **Parallel Production:** **OB52**
> Information for several stakeholders should be producible in parallel.
> **Prioritizability in Production:** The production of information for a stakeholder should be prioritized based on the urgency of the need. **OB53**

Furthermore, even without this problem, the explain system must be able to compute an information-timing with the largest possible benefit. But, in case an importance or urgency of the explanation is needed, it must be computable as well.

> **Computability of Benefit Maxima:** The information-timing with the highest benefit for a stakeholder should be computable. [22] **A54**
> **Computability of Need-Importance:** The importance of an explanation to a stakeholders need shall be computable [22]. **OA55**

Those requirements present the first for the analysis phase of MAB-EX.

5.1.7. Analytic and Computational Capabilities

In order to enable computations in general, the explain system should be able to access, use and analyze all the data it monitored and captured.

> **Analytic Capability:** **A56**
> The explain system should be capable of analysing data it observes.

Stakeholder

A data subset of particular interest is that of stakeholders. Already mentioned were the interests in their goals, their understanding of the explananda and the associated explanation needs. An additional advantage comes with the analysis of interactions between the whole system and stakeholders, in order to detect *aimless sequences* [9].

MAB-EX **Analysis of Goals:** The goals of stakeholders should be retrievable by analysing the interactions with, and behaviour of the system [9]. **A57**

MAB-EX **Analysis of Understanding:** A stakeholder's feedback should be analysable in order to verify if an information was understood. [9] **A58**

Computability of Need: A stakeholder's explanation need should be computable. **A59**

MAB-EX **Analysis of Interactions:** Interactions of the system and stakeholders should be analysable for aimlessness or contradictions [9]. **OA60**

Average and Individual Explanandum

To enable stakeholders to understand how individual elements of the context and their changes on average influence an explanandum, one can use the definitions of explainability-levels and their corresponding *meta-requirements* defined by Bersani et al. [7].[6]

Local Explanation: A local explanation for an individual explanandum shall be producible by considering all partial contexts. [7] **G61**

Global Explanation: A global explanation of the average of an explanandum shall be producible by considering all partial contexts. [7] **G62**

Thus, the context must be analysable into multiple contextual factors, and both an individual and an average explanation must be computable by calculating a sum of effects and an expected distribution, respectively.

MAB-EX **Factorization of Context:** The context shall be analysable as factors affecting the explanandum in the system's explainability goal. [7, 9] **A63**

Computability of Individual: An individual explanandum shall be computable as the sum of effects of the observable partial contexts. [7] **A64**

Computability of Average: The average of an explanandum shall be computable as the expected distribution of it. [7] **A65**

In addition, we need to require that the context is partially monitorable, and that measurements of the monitoring occur exactly when the explanations occur.

Partial Monitorability: Each partial context shall be monitorable by the corresponding agent participating. [7] **M66**

Behavioural Measurement-Availability: A measurement of the context-factors shall be available when the explananda occur. [7] **M67**

Thus this leads to the monitoring phase of MAB-EX, that deals with capturing the data needed for the analysis and calculations described above.

[6] An overview of the Meta-Requirements is in the Appendix.

5.1.8. Monitor and Data Capture

Before we look into this phase in detail, we need to require a final part of the system to be present, which is needed for all of the system's interactions with the context, including stakeholders: Interfaces. Among other things, these are needed for expressing needs, sensing the real world or other interactions of the system.

Presence of Interfaces: **P68**
The control system should have interfaces with the context.

The requirements of the monitoring phase can be divided into the capture of system specifics themselves and the observation of context entities.

System

The most important data needed to be captured by the explain system are interface events as previously motivated, sensor data, logs and commands used by the control system, and former explanations generated by it.

Capturability of Interface-Events: Possible interfaces of the control system must be observable by the explain system. **M69**

MAB-EX **Capturability of Sensor Data:** The explain system should be able to capture observer data used by the control system. [9] **M70**

MAB-EX **Capturability of Commands:** The explain system must be able to capture (former) commands used by the control system. [9] **M71**

MAB-EX **Capturability of Explanations:** The explain system should be able to capture (former) explanations. [1] **OM72**

Context

In addition, all context-entities present in the definitions of explanations and explainability must be monitorable, explicitly explananda, stakeholders and interactions with the system, but also any other context-detail of possible interest.

MAB-EX **Monitorability of Context:** The context shall be observable and measurable at runtime. [7, 9] **M73**

MAB-EX **Monitorability of Explananda:** The control system's explananda must be observable [7, 9] for stakeholders, and explain systems [9]. **M74**

MAB-EX **Monitorability of Stakeholder:** The explain system should be able to monitor a stakeholder including its behaviour and feedback. [9] **OM75**

MAB-EX **Monitorability of Interactions:** The explain system must be able to monitor interactions of the control system and stakeholders. [9] **M76**

Finally, the observed data should also be stored chronologically.

Chronological Capturability: The explain system should be able to capture the chronical aspects of all captured events. **OM77**

In summary, 77 theoretical architectural requirements were derived from the literature. Of those, about a third (25) were explicitly stated in papers regarding the release and development of the MABEX framework. The remaining requirements described ideas that had already been (partially) researched but not incorporated into the framework.

Figure 2 visualizes all requirements elicited. They are clustered according to the classifications defined at the beginning of Section 5 (e.g. according to their corresponding phase in MAB-EX). The four phase clusters are highlighted in purple, the explanation model cluster in teal. Requirements that need another to be fulfilled, or that are somehow linked with another aspect, are connected with an arrow (e.g. the 'comprehension of information' needs 'adaptability in its presentation').

—

This allows us to address Research Question 2, which asked *what requirements are needed for the implementation of self-explainable systems using MAB-EX.* Based on researched definitions of self-explainability and other helpful aspects for achieving such an ability in systems (timing, explanation models, etc.), all the theoretical requirements mentioned would be helpful for the implementation of them. While they may not all be necessary, fulfilling all the requirements would likely enable higher-quality explanations.

Answering Research Question 2: The previous 77 requirements were identified based on the current state of research and the literature we reviewed for this purpose. However, not all of these are strictly necessary for achieving self-explainable systems. Ideally, though, all of them should be fulfilled.

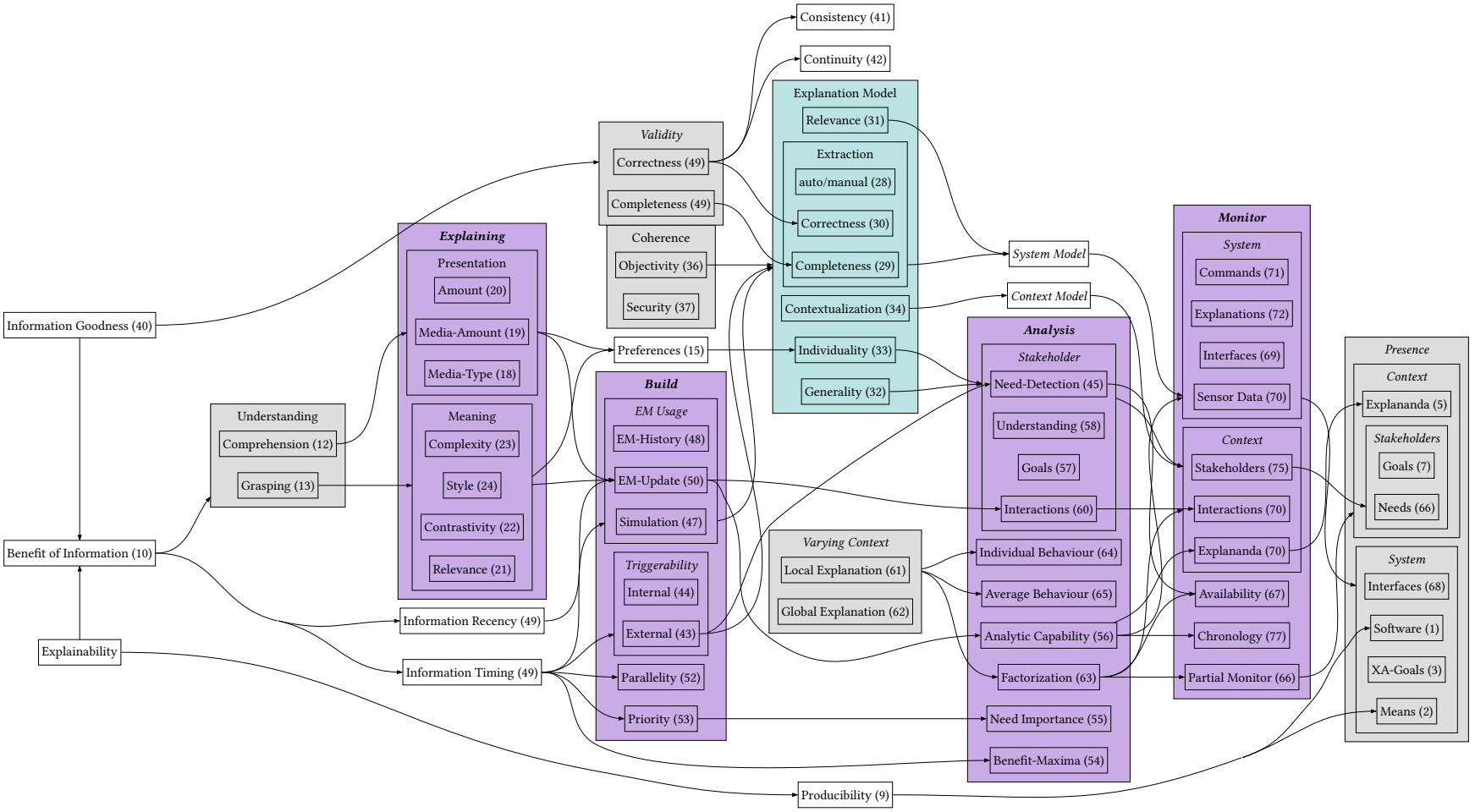

Figure 2: An overview on the requirements elicited; those that relate to each other are shown as clusters; clusters belonging to MAB-EX phases are highlighted in purple, the cluster that groups explanation model requirements is in teal.

6. Case-Study

Now that the theoretical requirements for self-explainable systems (SXS) have been established, their practical relevance will be determined through a case-study. For this, the six suggested practices for developing explainable systems by Chazette [10] will be followed. However, since the implementation it is not our objective, only a stakeholder-analysis (Practise 3) will be performed after defining the system's *vision* and *goals* (Practise 1). Practices (2) and (4-6) are out of scope for this thesis.[7]

6.1. System Description

The system is physically located at a traffic intersection. The environment is further composed of all things connected with road traffic, in particular road participants and the infrastructure required for traffic. An example setup can be simplified into Figure 3.

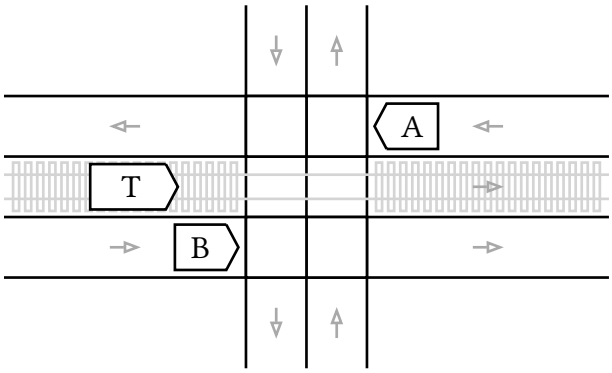

Figure 3: Example Scenario A: a Traffic Intersection; A, B and T are different vehicles; lane directions are marked with arrows.

Idea of the System

The idea of this Traffic Management System (TMS) is to dynamically manage and route traffic through the intersection in order to minimize waiting times for road participants. Because the decisions the system makes can be non-intuitive, or stakeholders could possibly have questions regarding phenomena in the systems environment, it should consist of two parts. One, the part managing the traffic, being the control system, and two an explain system part, that addresses explanation needs of stakeholders.

[7] Full practises are in the Appendix.

System Visions and Goals
Before requirements can be elicited, visions of the Traffic Management System should be phrased. A vision can refer to the capabilities, features, or quality aspects of a system [10], therefore being a close-to-reality concept of the desired future that describes what is to be achieved, but does not state how [29]. Thus, they are a reference starting point for the elicitation of requirements. The visions of the TMS can be described as:

Control System The control system should optimize the traffic flow with respect to its stakeholders needs. **V1**

Explain System The explain system should be able to explain the decisions and behaviour of the control system as well as the underlying traffic context accurate and correct. **V2**

For the purpose of this thesis, we will focus on goals and requirements regarding the explain system, excluding those of the control system in the following. Goals of a system can be seen as an intermediate format between visions and requirements. They are not as detailed as requirements, but at the same time not as general as visions. Goals regarding the explain system could be phrased as follows.[8]

Information Concept The primary purpose of the explain system is to explain the behaviour of the control system to traffic participants in order to help them understand the operational restrictions or possible consequences, and to create more trust in the system's decisions. **T1**

Information Benefit The information generated by the explain system should always benefit to goals of stakeholders using it. **T2**

Information Timing The explain system must be able to produce an explanation within an appropriate time of a stakeholder expressing an explanation need, dependent of the importance and urgency of the need. **T3**

Information Prediction In order to be able to generate explanations among others as fast as possible, the explain system shall be able to simulate possible future behaviour of the control system and the context. **T4**

Based on this, we can now begin analysing practical requirements, i.e. stakeholder needs, by eliciting explanation cases.

6.2. Basis for Requirements Elicitation

To meet the expectations of all stakeholders regarding explainability, Köhl et al. [4] proposed the following questions as a basis for elicitation of the requirements.

[8] Due to the same initials in 'Goals' and 'Global Requirements', the former are abbreviated as 'T'.

Table 3: Elicitation Questions adapted from Köhl et al. [4].

1. What are relevant target groups (G)?
2. What are the explananda (X)?
3. Which aspects (Y) of X must be explained to which G?
4. In which context (C) may an Y need an explanation?

Based on the resulting answers, we can then determine the explainability-requirements needed. As suggested by the authors of the questions, these requirements can mostly be understood as special non-functional requirements (NFR), and don't need to be satisfied, but rather *satisficed* [4]. The term satisfice is a fusion of satisfy and suffice, introduced by Simon [30], which is a good description of the nature of NFR, that often have to be traded off against each other when trying to fulfil them [4].

To follow the Questions in Table 3, different target groups are explored in Section 6.2.1, followed by an analysis of possible explananda in Section 6.2.2 and contexts in Section 6.2.3.

6.2.1. Possible Stakeholder / Target Groups

As an initial consideration, it is useful to differentiate between stakeholders who use the explain system in real time, i.e. parallel to events taking place, and those who use the system retro- or prospectively to gain insight into past or future events [5, 22]. To further subdivide the different explanation needs of real-time stakeholders, it is useful to divide them by their use of the underlying control system [5]. In the context of a Traffic Management System, it could be interesting to further categorize stakeholders based on their activity level and priority, depending on their level of urgency. For example, emergency vehicles could have the highest priority, while regular stakeholder, such as cyclists or drivers, could have the lowest. stakeholders could also be divided into privileged and non-privileged groups, such as system developers or operators, and public transport users. Finally, it could be of interest to categorize stakeholders who use the system for subsequent information into authorized and unauthorized groups. The former could include legal or government institutions, while the latter could include analysts, statisticians, researchers, or traffic engineers.

This means that the target groups in the following explanation-cases consist of ordinary road users, tram-operators and -passengers, emergency vehicle drivers, system-operators and -developers, researchers, and judges.

6.2.2. Possible Explananda

As stated in Section 2.1 possible explananda of a system can be the system in general, or a specific aspect of it, such as its reasoning processes, inner logic, model's internals, intention, behaviour, decision or knowledge [6]. In this case, possible behaviors based on the Traffic Management Systems decisions could include signal changes, (emergency)

warning signals, (dynamic) traffic prioritization or re-routing, and their underlying decision-making processes. Additionally, possible knowledge products could include (real-time) traffic reports, historical data, predicted trends, or informational notifications. These could also be represented at a technical level, e.g., as logs, models, or sensor states.

6.2.3. Contexts

In order to approach possible contexts of explanations systematically, example contexts were found for situation classes established by Sadeghi et al. [5]. They will be subsequently used in the cases. An overview of the examples compared to the their corresponding classes can be found in Table 4. For better comprehensibility, the overview is provided after the cases have been elicited, as these describe the example contexts most effectively. To avoid redundancy, an additional description has been omitted here.

6.3. Case Requirements Elicitation

In order to be able to establish requirements for the explain system based on the target groups that have emerged, we will phrase cases of individual stakeholders when encountering it. These will be presented as short stories and thus describe an explanation need of a single stakeholder in a specific context, following the syntax of explanation-cases used by Sadeghi et al. [5]. Also following a taxonomy, the same authors proposed, these cases can be divided into three categories: those that occur before, during or after a system's behaviour. Thus, they will be approached based on this distinction. To describe the stories in more detail, we will refer to Figure 3 by default, unless stated otherwise.

Note that the cases were selected based on the preceding taxonomy. In order to achieve complete coverage of stakeholder needs, it would be necessary to consider all possible cases. However, due to the scope of this thesis, that is not possible.

6.3.1. Ad-Hoc Explanations

We will start by analysing a stakeholder 'using' the system before an explanandum occurs. With this, the initial start of an interaction can lie on the side of the system, leading to this type of explanation.

S-1: Alice is on her way to work when she receives an information-notification from the TMS. The system informs her that due to construction on her usual route after the upcoming intersection, it would be faster to take a different route that day. Thus, her current lane is the slower one, and she should change lanes accordingly. The system decided to contact her before she would notice the problem herself so that she could save time and mental space on her way to work.

This story leads to multiple needs for self-explainable systems. First, the system must be able to detect deviations in context in order to process them into actions or explanations. The second requirement of this scenario is that the explain system has to be able to time its explanations in order to provide them before, during, or after phenomena of interest.

Deviation Detectability (based on S-1) **C1**
The system must be able to detect deviations in its context. Therefore, it must have the ability to observe and capture the context, as well as some understanding or representation of what counts as a deviation and what is normal. → confirms: P4, A63, M73

Timing of Explanations (based on S-1) **C2**
In order to maximize the benefit of an explanation for an addressee, the system must be able to give the information at the right time. For this, it must also be able to compute a point in time with maximized benefit, and to produce explanations in parallel. → confirms: G51, OB52, A54, M67

6.3.2. Real-Time Explanations

Apart from the previous case, the following continues with the analysis of stakeholders using the system in real-time. Most often, scenarios resulting in an explanation need are 'normal' real-time interactions with the system:

Interaction Cases

One of the most intuitive needs for an explanation in the traffic intersection scenario is a normal road user, such as a car driver, who, when faced with a new type of traffic scenario, may be confused and in need of an explanation of the situation.

S-2: There is normal, medium-volume traffic at the intersection. Bob is approaching the intersection for the first time in his non-autonomous vehicle. In order to be prepared for traffic situations, that he would not expect himself, he informs the TMS that he needs an explanation of the situation and common behaviors. Unfortunately, the system's explanation is too complicated for Bob to understand. The system detects this based on his facial expression. Therefore, it prepares a simpler version of the same content, and delivers it to him.

In this case, the explanation could just be a summary of the situation, an overview of the cause of it, or what potential consequences or dangers may arise. For instance, if a driver like Bob waits longer than expected at an intersection, it could be because an approaching vehicle has a higher priority, such as a tram. Based on Bob's initial lack of understanding,

this case yields the first user-requirement: the adaptability of the level of detail, i.e., the complexity of an explanation.

Level of Detail (based on S-2) **C3**
The system should be able to detect what level of detail is needed in a potential information for a target group. → confirms: OX23

In contrast to the previous case, Bob is familiar with the traffic situation this time. Therefore, his needs may have changed, but the context may not have. Nevertheless, he could request an explanation to make sure nothing has changed.

S-3: As Bob approaches the intersection again, he is already familiar with its characteristics. However, to make sure nothing has changed, he requests an explanation of the situation. The TMS provides a similar explanation because nothing in the context has changed. It also explains why the explanation hasn't changed. This strengthens Bob's confidence in the system and the upcoming situation.

Such consistency in production is desired to convey correctness to stakeholders through recognizability. This leads to a further requirement, that is relevant for the characteristics of explanation production.

Determinism (based on S-3) **NFC4**
The generation of information by the system should be consistent and continuous, i.e. if the context, including the target group, is the same, the resulting information must also be the same. Additionally, they should be similar, if the inputs were similar. → confirms: G41, OG42

Comparable to S-2, the opposite of a first-time encounter, a commuter passing by every day, might also experience a need when the system's behaviour deviates from the expected pattern. Although the cause may be different, the consequences could remain the same.

S-4: On a day, Alice is on her way to work when she approaches a intersection that is normally known to her. But today the traffic schedule seems to be off. To reassure herself that everything is under control, she contacts the TMS to get an explanation on the deviations. The system explains that an update to the routing algorithm has been implemented, so the routing doesn't behave the same way as before. Additionally, it explains why it didn't work as before and why the system chose not to make other routings.

However, based on this story, although the explanation need hasn't changed, this one provides a useful feature of explanations: it explains not only why an explanandum happened, but also why another didn't. Including partial explanations that describe the absence of another phenomenon requires the system to consider several variable contexts to account for different probabilities of occurrence of phenomena.

Contrast (based on S-4, S-7) **C5**
In order to let a representative of a target group understand the reasons to *why* an explanandum (X) happened, information on the contrast of X, e.g. *why not* another X happened, should be included. For this, *varying contexts* should be considered. → confirms: OX22, G61, G62, A65, M66, M74

An increasing prioritization in traffic goes hand in hand with an increasing urgency in the participant's progress in traffic. As a result, the main focus may change from the nature of the matter to an explanation that is available as quickly as possible. In the case of a tram operator as a stakeholder in need, this urgency is not very high but nevertheless important, as a timetable must be adhered to and passengers must be informed of delays.

S-5: It is rush hour. Tram operator Tim is approaching an intersection on schedule, but the signal from the Traffic Management System ahead tells him to stop and wait, which is outside the scheduled timetable. Tim requests information from the system to inform passengers of the reason for the stop and an estimate of the expected delay.

Apart from demonstrating the increased need and higher urgency, this story reveals another requirement for the system: In order to incorporate an estimation of the delay into the explanation, the explain system must be able to *predict* upcoming behaviour in order to make the estimation.

Simulatability of Behaviour (based on S-1, S-5) **C6**
In order to provide information about the consequences of deviations, the explain system should be able to simulate the future behaviour of the context. Therefore it needs a time or state model. → confirms: B25, B46 , B47

Even more so than with S-5, urgency takes on a significant role at the highest prioritization level. For example, an emergency vehicle driver is interested in crossing an intersection as quickly and safely as possible. It should therefore be possible to generate an explanation quickly, but still with all the necessary information.

S-6: On her way to the hospital, the emergency vehicle driver Eve approaches an intersection where the traffic situation is not visible from a distance. So she contacts the Traffic Management System to get a quick and reliable explanation of how to get through the junction as fast and safe as possible. In order to save time and mental space, the Traffic Management System generates the information as visual into the routing service of the vehicle.

The requirements for the explain system resulting from this story are highly relevant because time and validity can be crucial in such situations. First, similar to how the emergency vehicle is prioritized in the traffic routing of the control system, the need for an explanation should be prioritized in the production of it by the explain system. To this end, the explain system must also be able to compute the urgency of the need.

Prioritizability of Generation (based on S-6) **C7**
When interacting with multiple target groups, the system must analyze their priority to ensure that those with the greatest need receive their necessary information first. → confirms: OB53

Urgency of Information-Need (based on S-6) **C8**
In order to maximize the benefit of the information generated, the system must be able to compute the urgency of the target group's need. → confirms: OA55

As a passive road participant, the need for an explanation from the perspective of a tram passenger opposed to that of a tram operator (S-5) is in principle very similar, but technically less focused on the details and more on a short, concise explanation including the resulting consequences.

S-7: To get to the main station, Patrick takes the tram as a passenger to get there on time, as he is already late. When the tram slows down at an intersection where it normally doesn't, he wonders if the tram will be delayed and if he will have to book another connecting train. The Traffic Management System tells him with a short information-notification that the tram is on time and has just slowed down because it is ahead of schedule.

In this story, the priorities of the stakeholders are not as important as the way the explanation is presented to it. Based on its needs, preferences, and knowledge of the world, the presentation that benefits a holder the most can vary significantly.

Representation of Content (based on S-6, S-7) **C9**
The information generated by the system must be adaptable in its presentation. Hence, the system must be able to identify the most appropriate way to present information to a target group. → confirms: X16, OX18, OX19, OX20, OX24

Another aspect showcased is that the stakeholder did not explicitly phrase their need to the system, but rather, it detected its confusion and generated an explanation to address it. This results in the requirement that the explanation production should be triggerable by both the system or a stakeholder in need.

Triggerability of Generation (based on S-7) **C10**
The generation of information by the system must be triggerable externally through the explicit expression of a target group or internally when an implicit explanation need is detected. Thus it must also have an interface where a need can be detected. → confirms: B43, B44, A45, A59, P68, M69

Debugging Cases

Another type of explanations can be described as *debugging*, in which the interaction of a stakeholder with a system focuses on analysing the system's behaviour for malfunctions rather than the intended use of it. These details may be of interest to a system-operator or a software developer in order to avoid possible causes of problems in the future.

S-8: Ollie, the system operator in charge, monitors the Traffic Management System in real time, looking for alerts about anomalies in the current behaviour. Thus, he wants detailed information about the underlying roots of its decisions and behaviors to ensure the performance and functionality of the system so that he can intervene if necessary.

If, for example, the control system has made a wrong decision in the prioritization of participants from the operator's point of view, regardless of the reason, the operator should be able to intervene as precisely as possible, and then needs details about the cause of the decision. This requires that the system can accurately reproduce the correlations between events and its behaviour.

Correlations of Events (based on S-8, S-9, S-10, S-11, S-12) **C11**
In order to generate explanation for certain behaviour, the system must be able to detect and capture correlations between events in the context and resulting ones of the control system. → confirms: B25, B26, E34

Even more detailed, and possibly not in natural language, behaviour and decisions could even be of interest to a system developer at the highest level, in order to be able to adapt a behaviour not temporarily, but in a permanent way.

S-9: To adjust the behaviour and decisions of the Traffic Management System and ensure that they are in line with the company's desired policies, software engineer Sarah requests system logs, generated models and the resulting behaviour, as well as the underlying correlations between them, in as much detail as possible. To satisfy this request, the system generates the desired information in text form to explain complex relationships in detail. Somehow, Sarah does not seem to understand the generated information at first. The system notices this, re-evaluates, and generates another explanation, this time also using visuals.

Apart from requirement C11, which is the correlation between events, this case also highlights that events and their relations should be storable in the correct historical order, which can be used to understand explananda. This results in another requirement:

History of Behaviour (based on S-9, S-10, S-11, S-12) **C12**
In order to generate an explanation of the control system's past decisions and behaviour, it must be able to capture and store relevant data (logs, models, explanations) in the correct historical order. → confirms: E48, M70, M71, OM72, OM77

For stakeholders, who are subsequently interested in explananda of the control system, completely different details may be of importance.

6.3.3. Post-Hoc Explanations

In contrast to active holders, a researcher, for example, might neither be interested in quick nor short explanation of behaviour, but rather in the interactions within the control system. This could include the correlations between context events and system behaviour based on these events.

Research

As the data examined in such a case could also be used as the basis for a scientific study, the system should only provide it under protective measures. These could, for example, be motivated by data privacy.

S-10: As a researcher, Robert is also interested in the relationships between the behaviour and decisions of the Traffic Management System. To this end, he requests such data from the system. The information provided by the system is provided in anonymized form because the system cannot guarantee that the requested data will not be used for publication, e.g. in a scientific paper.

Furthermore, this holder could also be interested in detailed log-like structures in order to improve a possible behaviour prediction model. Also, government officials or similarly authorized holders should also possibly have more permissions regarding the use of data for analysis, provided that the use remains compliant with data protection laws. For example, data sets could also be used for the purposes of making the traffic flow more efficient and to enable (legal) changes in the prioritization of traffic participants.

S-11: In addition, to train the predictive capabilities of the Traffic Management System, Robert needs log-like behavioural structures as a training set. For this purpose, he requests the data with a special access allowing him to obtain the data in raw form. For this job, he was hired by the local government to optimize the traffic flow with a possible more efficient model, which required him to sign a contract guaranteeing the privacy of the collected data.

As mentioned, these cases highlight the importance of data security. This is not only for legal reasons, but also to ensure stakeholders interact with the system with trust.

Data Privacy/Security (based on S-10, S-11, S-12) **NFC13**
The sensitive data collected and stored by the explain system must always be handled under strict privacy and security measures in order to ensure stakeholder trust in the system. → confirms: E37

Validation

In certain cases, however, authorized stakeholders should be able to access any data retrospectively. For example, law enforcement groups should be able to check the system for compliance with laws or regulations, for instance to check the behaviour of the system for potential misbehaviour following a traffic accident.

S-12: In order to rule on a case involving an accident on the site of the Traffic Management System, Judy, the judge, needs to check how the system has reacted and whether its decisions and actions were in accordance with the law. For this reason, she needs all the data generated in a form that is verifiably correct, so that she can use it as legal evidence.

The relevant aspect of this story is that the information transmitted by the system must be reliable and valid. Because if the system cannot guarantee that, then explanations cannot build trust into decisions, and especially cannot be used as evidence.

Validity of Information (based on S-6) **NFC14**
The information generated by the system must be provably valid in order to be beneficial for the target group and to build trust in the system. Thus, also the information used as a basis for build models must also be valid. → confirms: OE29, OE30, G38, G39

We elicited requirements based on specific cases to highlight their origin. However, some of the characteristics required for an self-explainable system are based on all of the aforementioned cases.

Additional Requirements based on all Cases

In all cases, an explanation need was addressed and fulfilled by transmitting information. This takes us back to the general definition of explainable systems. As discussed in more detail in the previous theoretical elicitation, this also leads to information needing to be producible in some way, by some means.

Explainability (based on *all cases*) **NFC15**
The system must be explainable for a target group in a context with respect to an aspect of an explanandum [4]. → confirms: P5, M74, OM75

Producibility of Information (based on *all cases*) **C16**
The system should be able to produce some information in order to address potential needs of target groups. Therefore it has to have some means. → confirms: P2, X9

As many cases involved stakeholders interacting with the system in real time and requiring a quick explanation, information has to be produced in real time.

Real-Time Generation (based on *all cases*) **C17**
The information required by target groups must be generated at runtime by the system in order to be up-to-date. For this reason, also prior system interactions with target groups must be taken into account in potential future information-generation. → confirms: B49

A comparison of all previous cases also makes it clear that the underlying needs and preferences of stakeholders for explanations were not always the same. This confirms the need for information to be customizable.

Individuality of Information (based on *all cases*) **C18**
To maximize the benefit of an explanation for an addressee (A), the system should adapt the information to align with A's current (world) knowledge. The information should be coherent with A's knowledge, but should not contain any bias towards A, and therefore be objective. → confirms: G14, B15, G35, G36, A56, A64

At the same time, the information produced should also be of use to the stakeholders.

Both S-2 and S-9 demonstrate that this often necessitates the exchange of multiple pieces of information. In order to recognize the understanding of an explanation, it could be sensible to introduce a threshold above which understanding is considered sufficient. This also means that stakeholder-understanding has to be measurable in some way.

Benefit of Information (based on *all cases*) **C19**
Through receiving an explanation, the distance to a threshold representing the understanding of an explanandum shall decrease in order for the explanation to be beneficial to the addressee's understanding of the explanandum. → confirms: G10

Measurability of Understanding (based on *all cases*) **C20**
In order to measure whether the needs of a target group were satisficed or whether further explanations are needed, the system should be able to evaluate whether the target group has understood the information provided. → confirms: G11, G12, G13, A58

Given the importance of certain explanation needs in these cases, the production of information regarding an explanandum should only include relevant aspects to it.

Relevance of Information (based on *all cases*) **C21**
The information produced by the system should be of relevance in order to be beneficial to the goals of both entities. Thus, the system must be able to analyze the addressee's goals or motivations and (former) interactions between them. → confirms: OX21, A57, OA60, M76

As mentioned at the start of Section 6.2.3, the cases were approached systematically by following the taxonomy proposed by Sadeghi et al. [5]. They were differentiated by the timing of the explanation prior, during or after an explanandum, and the type of explanation situation.

Table 4: Taxonomy of Explanation Contexts and their Timing by Sadeghi et al. [5] with Example Contexts resulting in Explanation Needs.

Taxonomy (Types of Situation)			**Time**	**Example Contexts**
training			prior	road participants first time encounter
interaction	disobedience	goal-order conflict	during	no priority given to faster or easier route
		multi-user conflict		question of right of way
		contextual condition		request not fulfilled due to other priority
		system condition		*no condition to be met necessarily for TMS*
	failure	system error		malfunctioning signals
		user fault		deviation in stated and actual user action
	context-aware behavior	suggestion		re-routing due to deviation in traffic flow
		autonomous action		dynamic reroute based on traffic volume
debugging			after	(software) engineer maintaining system
validation				historical evidence for legal entities

—

In summary, this chapter has identified 21 practical user-requirements. Due to the user-perspective, these requirements apply to the entire system rather than to individual components, so they are mostly non-technical. Consequently, they tend to describe both broad functionalities and desired Non-Functional Requirements. As such, these requirements often inspire ideas that were previously described in theoretical requirements through several more detailed requirements. Therefore, they confirm the use of the theoretical requirements in the potential framework.

7. Evaluation

To evaluate the artefacts identified so far in the form of requirements from theoretical and practical perspectives, specifically in terms of architectural- and user-perspectives, we will compare the coverage of both artefact sets below.

Note that more than three times as many requirements were found in the theoretical analysis than in the practical analysis. This is due to the fact that they are based on a much larger theoretical foundation that has already been elaborated upon, and therefore could be expressed in much more detail. Those from the practical case study, on the other hand, come from the users' point of view and therefore describe much coarser requirements for the system.

To address the question of *whether the integrity of MAB-EX can be validated by comparing explainability-requirements from theory and practice*, we will split the comparison based on their classification made in Section 5.

Additionally, to determine whether case-requirements are covered by theory requirements or vice versa, we will analyze these two aspects separately. First, we will analyze the former, and then the latter. To highlight requirements that have been addressed and required by publications referring to MAB-EX, these are stated in bold with (MAB-EX) in the following, because they are of special interest regarding the research question. Key points derived from the comparison are enumerated to refer to them later.

7.1. Confirmation of Architectural-Requirements through User-Requirements

7.1.1. Presence Requirements

The presence requirements explicitly required by MAB-EX are those of a context, phenomena, and an explainability goal. As seen in Table 5 the first two can be confirmed directly through case-requirements; the latter cannot. The rest of the theory requirements that are not explicitly needed by MAB-EX are either confirmed by case-requirement or implied as well.

Table 5: Confirmation of Presence-Requirements by Case-Requirements.

Presence-Requirement	ref	Confirmed through
Presence of Software System	P1	*implied by* Section 6.1
Presence of Means	P2	C16
Presence of Explainability Goal MAB-EX	P3	*implied by Section 6.1*
Presence of Context MAB-EX	P4	C1, *implied by* Section 6.1
Presence of Phenomena MAB-EX	P5	NFC15
Presence of Stakeholders	P6	*implied by* Section 6.2.1
Presence of Stakeholder Goal	OP7	*implied by all stories*
Presence of Explanation Needs	P8	*implied by all stories*
Presence of Interfaces	P68	C10

For example, the case-requirement 'Triggerability of Generation' C10, covers the theory requirement 'Presence of Interfaces' P68 because, in order to be triggerable from outside, the system needs interfaces with the environment.

Remark 1: *All presence requirements are either confirmed directly or implied indirectly.*

7.1.2. General Requirements

Continuing with the general requirements from Section 5 in Table 6, it might be confusing that none of them are marked as required by MAB-EX. However, this is because none of them are *explicitly* stated as required in any of the papers regarding the framework.

Schwammberger et al. [1] argue that MAB-EX already supports the idea of local G61 and global G62 explanations, and suggest adding support for multiple agents to align with the idea of explainability-levels. Schwammberger [3] also suggests explicitly incorporating the idea of explanation goodness G40 and correctness G38.

Table 6: Confirmation of General-Requirements by Case-Requirements.

General-Requirement	ref	Confirmed through
Benefit of Content	G10	C19
Understandability of Content	G11	C20
Comprehensibility of Content	G12	C20
Grasping of Content	G13	C20
Adaptability of Information	G14	C18
Information Coherence	G35	C18
Information Objectivity	G36	C18

Information Correctness	G38	NFC14
Information Completeness	G39	NFC14
Information Goodness	G40	*transitive by* G11, G38
Information Consistency	G41	NFC4
Information Continuity	OG42	NFC4
Information Timing	G51	C2
Local Explanation	G61	C5
Global Explanation	G62	C5

Remark 2: *All general requirements are confirmed by case-requirements. (Regardless of whether they are explicitly required by MAB-EX)*

This is especially interesting because these requirements pose the greatest challenges to self-explainable systems and, consequently, to a framework designed to build them.

7.1.3. Explaining-Phase Requirements

A similar picture emerges when examining the coverage of the requirements from the explanation phase of MAB-EX:

Remark 3: *All explanation requirements are confirmed by case-requirements.*

Table 7: Confirmation of Explanation-Requirements by Case-Requirements.

Explanation-Requirement	ref	**Confirmed through**
Producibility of Content MAB-EX	X9	C16
Adaptability of Presentation Syntax MAB-EX	X16	C9
Adaptability of Presentation Semantics	X17	*transitive by* C9
Adaptability of Media	OX18	C9
Adaptability in Media-Amount	OX19	C9
Adaptability in Amount	OX20	C9
Adaptability of Relevance	OX21	C21
Adaptability of Contrast	OX22	C5
Adaptability of Complexity	OX23	C3
Adaptability of Style	OX24	C9

Note that tables comparing requirements linked to phases or the explanation model are highlighted in the same color used to identify them previously.

However, not all of them are marked as required by MAB-EX. This is because, so far, the papers related to the framework have addressed the presentation of information in general, but none have gone into more detail.

Remark 4: *Details on information-presentation could be used to refine MAB-EX.*

Thus, details on the aspect of information-presentation found in the analysis are not required by MAB-EX, but could be used as a refinement of it.

7.1.4. Build-Phase Requirements

At first glance, we see that not all of the requirements from the build phase are confirmed by case-requirements. However, those that are not confirmed are on the technical side, which explains why they are not confirmed by case-requirements, i.e. stakeholder needs.

Remark 5: *Some technical build requirements are not confirmed by case-requirements.*

These requirements, though, are linked to the idea of explanation models and are thus relevant to the usage of one, which confirms there relevance for MAB-EX.

Table 8: Confirmation of Build-Requirements by Case-Requirements.

Build-Requirement	ref	**Confirmed through**
Preferences of Stakeholder	B15	C18
Buildability of Explanation-Model MAB-EX	B25	C6, C11
EM Initializability	B26	C11
External Triggerability MAB-EX	B43	C10
Internal Triggerability MAB-EX	B44	C10
Extractability from Explanation Model MAB-EX	B46	C6
Simulation of Behaviour MAB-EX	B47	C6
Information Recency	B49	C17
Updatability of Explanation Model	B50	
Parallel Production	OB52	C2
Prioritizability in Production	OB53	C7

The rest of the requirements that are not linked to the explanation model in this phase are confirmed directly through case-requirements. This means that every build requirement explicitly required by MAB-EX is also confirmed, especially the ‘Simulatability of Behaviour’ C6, which is necessary to provide explanations at the right time.

Remark 6: *All build requirements required by MAB-EX are confirmed by case-requirements.*

7.1.5. Requirements for Explanation Models

The requirements for an explanation model are similar to those for the build phase, but not all are confirmed due to their technical aspects.

Table 9: Confirmation of Explanation Model-Requirements by Case-Requirements.

Explanation Model-Requirement	ref	Confirmed through
Manual Constructability MAB-EX	OE27	
Automated Constructability MAB-EX	OE28	
Complete Constructability	OE29	NFC14
Correct Constructability	OE30	NFC14
EM Relevance	E31	
EM Generalizability	E32	
EM Individuality	E33	
EM Contextualization	E34	C11
Information Security	E37	NFC13
History of Explanation Model MAB-EX	E48	C12

For the same reason, the only ones required directly by MAB-EX are not confirmed either, since they only require information about how the explanation model is built before deployment, which is not useful to stakeholders using the system in real-time. Taking into account that some of the requirements on explanation models are based on the work of Schwammberger et al. [2], which could also be considered an adaptation to MAB-EX [2], these requirements could also be referred to as 'MAB-EX requirements'. With this, even more requirements needed by MAB-EX with the explanation model wouldn't be confirmed by case-requirements.

Remark 7: *Some technical explanation model requirements are not confirmed by case-requirements.*

7.1.6. Analysis- and Monitor-Phase Requirements

This section will examine the requirements coverage of the analysis and monitoring phases of MAB-EX together. This is because they are closely linked and necessary for implementing more abstract requirements or preparing others in later phases. For instance, to trigger information-production in the build phase, the system must first detect the need for explanations in the analysis phase. This requires monitoring the relevant stakeholder-data in the monitor phase. (B44 → confirms: A45 → confirms: OM75)

Table 10: Confirmation of Analysis- and Monitor-Requirements by Case-Requirements.

Analysis- and Monitor-Requirement	ref	**Confirmed through**
Detectability of Explanation Needs MAB-EX	A45	C10
Computability of Benefit Maxima	A54	C2
Computability of Need-Importance	OA55	C8
Analytic Capability	A56	C18, *transitive by* A57
Analysis of Goals MAB-EX	A57	C21
Analysis of Understanding MAB-EX	A58	C20
Computability of Need	A59	C10
Analysis of Interactions MAB-EX	OA60	C21
Factorization of Context MAB-EX	A63	C1
Computability of Individual	A64	C18
Computability of Average	A65	C5
Partial Monitorability	M66	C5
Behavioural Measurement-Availability	M67	C2
Capturability of Interface-Events	M69	C10
Capturability of Sensor Data MAB-EX	M70	C12
Capturability of Commands MAB-EX	M71	C12
Capturability of Explanations MAB-EX	OM72	C12
Monitorability of Context MAB-EX	M73	C1
Monitorability of Explananda MAB-EX	M74	C5, NFC15
Monitorability of Stakeholder MAB-EX	OM75	NFC15
Monitorability of Interactions MAB-EX	M76	C21
Chronological Capturability	OM77	C12

This means that all these necessary technical requirements are confirmed by case-requirements that also cover the corresponding abstract requirements, which describe the functionalities achieved by implementing the technical ones. Consequently, all monitor- and analysis-requirements specified by MAB-EX are confirmed by case-requirements.

Remark 8: *All monitor- and analysis-requirements required are confirmed by user needs.*

This also means that, compared to the other phases, these two phases motivated by MAB-EX are already quite well defined.

7.2. Coverage of User-Requirements through Architectural-Requirements

Considering the coverage from the perspective of whether the case-requirements are covered by at least one of the theory requirements reveals that they are, as shown in Table 11. This suggests that the research is headed in the right direction.

Remark 9: *Each case-requirement is covered by at least one theory requirement.*

However, the outcome is different if we only consider those required by MAB-EX. Some are not covered at all, while others are only partially covered by presence- or monitoring-requirements.

Remark 10: *Not all case-requirements are covered by theory requirements that are explicitly stated in a paper regarding the development of MAB-EX.*

To understand this in more detail, we will examine each affected requirement individually in the following.

Table 11: Coverage of Case- through Theory-Requirements; MAB-EX Requirements are highlighted in bold.

Case-Requirements	ref	**Modelled by**
Deviation Detectability	C1	**P4**, **A63**, **M73**
Timing of Explanations	C2	G51, OB52, A54, M67
Level of Detail	C3	OX23
Determinism	NFC4	G41, OG42
Contrast	C5	OX22, G61, G62, A65, M66, **M74**
Simulatability of Behaviour	C6	**B25**, **B46**, **B47**
Prioritizability of Generation	C7	OB53
Urgency of Information-Need	C8	OA55
Representation of Content	C9	**X16**, OX18, OX19, OX20, OX24
Triggerability of Generation	C10	**B43**, **B44**, **A45**, A59, P68, M69
Correlations of Events	C11	**B25**, B26, E34
History of Behaviour	C12	**E48**, **M70**, **M71**, **OM72**, OM77
Data Privacy/Security	NFC13	E37
Validity of Information	NFC14	OE29, OE30, G38, G39
Explainability	NFC15	**P5**, **M74**, **OM75**
Producibility of Information	C16	P2, **X9**
Real-Time Generation	C17	B49

Individuality of Information	C18	G14, B15, G35, G36, A56, A64
Benefit of Information	C19	G10
Measurability of Understanding	C20	G11, G12, G13, **A58**
Relevance of Information	C21	OX21, **A57**, **OA60**, **M76**

The requirement 'Level of Detail' C3 is not directly covered by MAB-EX's requirements, but arguably indirectly by the more general requirement 'Representation of Content' C9, since complexity of information can be interpreted as a type of representation.

Similarly, the requirement of 'Contrast' in information C5 can be interpreted as a type of adaptability in presentation. Currently, MAB-EX's theory requirements only partially cover this. However, if the ability to include variable contexts in explanations, developed in Bersani et al. [7], is also implemented in MAB-EX, as described in Schwammberger et al. [1], then this requirement would be more distinctly covered.

Furthermore, the 'Determinism' requirement NFC4 would also be covered by the MAB-EX requirements if the findings from Schwammberger [3] were already part of it, as previously mentioned. They propose requirements of correctness and deducibility of explanations, which, when combined, can be considered equivalent to a deterministic production.

Similarly, it could be argued that the 'Real-time Generation' requirement C17 could be covered by incorporating insights from Schwammberger et al. [2], as they propose a process requiring an explanation model to be updated at runtime and information to be extracted from it in a similar way. By incorporating their proposed process, information would be adapted to individual stakeholders, thus covering the requirement of 'Individuality of Information' in more detail, which has only been partially covered so far.

Assuming that both mentioned papers would also be incorporated into MAB-EX, the requirement 'Validity of Information' NFC14 could also be considered covered. This could be inferred as a combination of the correctness of the information required in the first paper and the completeness of the information needed in the system model proposed in the second paper.

By incorporating the insights from Bairy et al. [22] into MAB-EX, the concept of 'Justifications' proposed there could also address the requirement for 'Validity of Information' NFC14. Furthermore, important insights regarding the timing of explanations are presented there, which, if included, would also cover the case-requirement of the same name C2.

In summary, it would be beneficial for the quality of explanations to incorporate insights from Schwammberger [3], Schwammberger et al. [2] and Bairy et al. [22] into MAB-EX, as described previously.

Remark 11: *It might be beneficial to incorporate the proposed insights on explanation-timing, -justification, -correctness and -models into MAB-EX.*

Three other requirements that were not directly covered are the 'Prioritizability of Generation' C7, the 'Urgency of Information-Need' C8 and 'Data Privacy/Security' NFC13. The latter has also been identified as an important requirement in other studies. However, to the best of our knowledge, no approach has yet been proposed for implementing these three requirements in self-explainable systems.

Remark 12: *Aspects connected to data security and the priority in production might be needed in MAB-EX.*

One final requirement that was not covered by the comparison is the 'Benefit of Information' C19: This is accompanied by the case-requirement 'Measurability of Understanding' C20, which is only partially covered by the theory requirement 'Analysis of Understanding' A58, and does not fully address the need described in the case-requirement. As previously mentioned in Section 5.1.2, the fact that the information conveyed by the explain system must be genuinely understood by the intended stakeholder is crucial for explainability.

The division of this point — apart from the comprehensibility G12 — into the 'Understandability of Information' G11 and the aspect of whether the information has reached the explainability goal illustrates the problem: while logical understanding may be measurable by techniques such as analysing the stakeholder's facial expressions, it does not necessarily indicate whether the explanation is sufficient.

To achieve the latter, it would be necessary to obtain feedback from the stakeholder. This was also suggested by Bersani et al. [7], who demanded that explanations exceed an explanation quality value chosen by the stakeholder. Consequently, this value should be calculable, or manual feedback must be possible. Schwammberger et al. [1] took up and refined this idea, though not to the extent necessary for incorporation into MAB-EX.

Thus, incorporating an explanation quality as proposed and refined would be needed for the goal of self-explainability, ensuring stakeholder-understanding.

Remark 13: *The idea of an explanation quality incorporated into MAB-EX is beneficial for self-explainability, supporting the understanding by stakeholders.*

7.3. Discussion

With the found remarks[9], we can now try to answer Research Question 1, asking whether the MAB-EX Framework can be validated by comparing these requirements.

Combined, Remarks 1 – 3 and 5 – 8 state that, apart from some technical requirements, all theoretical requirements can be confirmed by practical requirements. This means that

[9] An overview is in the Appendix.

all architectural requirements for self-explainable systems can be met from the point of view of requirements derived through stakeholder-needs. This underlines what was also found in Remark 9: this confirmation suggests that Explainability Research is generally heading in the right direction.

Remark 10 stated that the restriction of theory requirements to those explicitly addressed in research related to MAB-EX leads to not all case-requirements being covered. Combining this with Remarks 4 and 11 – 13, which provided examples of how MAB-EX could be carried out using both researched and new aspects, raises the question of how detailed MAB-EX should be specified, or if it, being a reference framework [9], should remain a coarser blueprint.

However, in order to answer this question, the level of detail that a reference framework needs to have must be answered. If MAB-EX is intended to be a general 'reference model', the level of detail should be limited to a minimal set of concepts independent of concrete details [19]. However, if the intention is for it to be a framework containing all the elements and phases necessary for implementing a solution [20], additional details might need to be incorporated.

With this, we can now give a dependent answer to Research Question 1.

> **Answering Research Question 1:** (A) The benefit of the current state of MAB-EX can be validated from stakeholder-side. (B) If MAB-EX aims to be a complete framework for self-explainable systems, more details might need to be incorporated.

7.4. Limitations

Coverage of Cases Due to the small number of cases in the analysis of requirements from the stakeholder's perspective, it is possible that not all aspects that could be placed on self-explainable systems were observed. However, since the coverage aligns fairly well with the current state of research, from our perspective, at least, nothing obvious has been overlooked.

Research-Scope Building on the above, it is possible that important literature was overlooked or not found when analysing the theoretical requirements for self-explainable systems. One could conduct this research more systematically in a SLR to eliminate this uncertainty. However, literature research was not the main scope of this thesis.

Level of Detail As we found, the answer to Research Question 1 depends on your demands on a framework. Depending on these demands, the requirements may not have been worked out in sufficient detail or may be too detailed. Accordingly, the research question was only answered in a dependent manner.

8. Conclusion

8.1. Summary

The goal of this thesis was to identify requirements on self-explainable systems (SXS) from theoretical (architectural) and practical (user) perspectives (RQ 2). Through a comparison of these requirements, an analysis was conducted to determine whether the current state of the MAB-EX-Framework is sufficient as a reference framework for building self-explainable systems or if it lacks aspects that should be included (RQ 1).

To this end, a literature review was conducted to examine the current state of MAB-EX and related research in Explainability Engineering necessary for developing self-explainable systems. To compare the resulting theoretical requirements (TR) with practical requirements, a case study was conducted on a Traffic Management System (TMS). These case-requirements (CR) were identified by conducting a stakeholder-analysis with specific explanation-cases to determine needs onto the TMS.

Through the literature-review, 77 architectural theory requirements were identified that are needed for self-explainable systems (SXS), thus being related to MAB-EX's goal of enabling systems to achieve this. To answer RQ 2, not all of them are strictly necessary for achieving self-explainability, but, ideally, all should be met. In the TMS case-study, 21 more general user case-requirements were found. Unlike those found in the requirements analysis, these did not describe concrete implementations or technical details, but rather abstract functionalities the TMS, and with this SXS in general, would need.

In the comparison of coverage of both requirements-sets, some case-requirements summarized what was found in multiple theory requirements. Apart from some theoretical requirements that were very technical, all could be confirmed by case-requirements. These bidirectional connections between practical and theoretical requirements meant, that all case-requirements were also covered by theory requirements. However, only considering theoretical requirements that were explicitly incorporated into MAB-EX to date, not every case-requirement was covered.

This can be interpreted as confirmation that Explainability Research towards SXS is generally heading in the right direction. But, the question arises how detailed MAB-EX, as a reference framework, should be. To answer RQ 1, we propose the following:

(A) The benefit of the current state of MAB-EX can be validated from stakeholder-side;
(B) If MAB-EX aims at being a complete framework for self-explainable systems, it may require more details to be incorporated.

8.2. Future Work

The elicited theoretical requirements can be seen and used as a knowledge base to further investigate the development of MAB-EX or other topics regarding the development of self-explainable systems.

MAB-EX Refinement Including key aspects that have not been fully covered in MAB-EX research into the framework could highlight further aspects in the field of self-explainable systems. This could include topics such as explanation-timing, -justification, -correctness, -quality and -models, as well as privacy, security and priority.

More Detailed Requirements To elaborate further on the architectural requirements for self-explainable systems, one could include papers that provide more detail on specific requirements. For example, these papers could address aspects such as explanation quality, mental models, and explanation validity.

Uniform Explainability Definition Another topic of interest for further research is establishing a standardized definition of explanations and explainability. Currently, there are different definitions, which often lack important aspects. It is important to determine which aspects should be included in a uniform definition since it also determines which functionalities self-explainable systems must be able to fulfil.

More Explanation-Cases To better validate the identified architectural requirements from users' perspectives, exploring more explanatory cases in several different scenarios could provide more insight that could be useful for developing self-explainable systems in general and MAB-EX in particular.

Bibliography

[1] M. Schwammberger, R. Mirandola, and N. Wenninghoff, "Explainability Engineering Challenges: Connecting Explainability Levels to Run-time Explainability," in *Proceedings of 2nd World Conference on Explainable Artificial Intelligence Conference (XAI2024)*, 2024. doi: https://doi.org/10.1007/978-3-031-63803-9_11.

[2] M. Schwammberger and V. Klös, "From Specification Models to Explanation Models: An Extraction and Refinement Process for Timed Automata," in *FMAS/ASYDE@SEFM*, 2022. doi: https://doi.org/10.48550/arXiv.2209.14034.

[3] M. Schwammberger, "From Explanation Correctness to Explanation Goodness: Only Provably Correct Explanations Can Save the World," in *AISoLA*, 2024. doi: https://doi.org/10.1007/978-3-031-73741-1_19.

[4] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender, "Explainability as a Non-Functional Requirement," *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pp. 363–368, 2019, doi: https://doi.org/10.1109/RE.2019.00046.

[5] M. Sadeghi, V. Klös, and A. Vogelsang, "Cases for Explainable Software Systems: Characteristics and Examples," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 181–187. doi: https://doi.org/10.1109/REW53955.2021.00033.

[6] L. Chazette, W. Brunotte, and T. Speith, "Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue," *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pp. 197–208, 2021, doi: https://doi.org/10.1109/RE51729.2021.00025.

[7] M. M. Bersani, M. Camilli, L. Lestingi, R. Mirandola, M. G. Rossi, and P. Scandurra, "A Conceptual Framework for Explainability Requirements in Software-Intensive Systems," *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pp. 309–315, 2023, doi: https://doi.org/10.1109/REW57809.2023.00059.

[8] L. Chazette, "Requirements Engineering für Erklärbare Systeme," in *Ausgezeichnete Informatikdissertationen 2022 (Band D23)*, Gesellschaft für Informatik e.V., 2023, pp. 31–40. doi: https://dl.gi.de/handle/20.500.12116/42607.

[9] M. Blumreiter *et al.*, "Towards Self-Explainable Cyber-Physical Systems," *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and*

Systems Companion (MODELS-C), pp. 543–548, 2019, doi: https://doi.org/10.1109/MODELS-C.2019.00084.

[10] L. Chazette, "Requirements engineering for explainable systems," Institutionelles Repositorium der Leibniz Universität Hannover, 2023. doi: https://doi.org/10.15488/13261.

[11] M. C. Buiten, L. A. Dennis, and M. Schwammberger, "A Vision on What Explanations of Autonomous Systems are of Interest to Lawyers," *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pp. 332–336, 2023, doi: https://doi.org/10.1109/REW57809.2023.00062.

[12] W. Brunotte, L. Chazette, V. Klös, and T. Speith, "Quo Vadis, Explainability? - A Research Roadmap for Explainability Engineering," in *Requirements Engineering: Foundation for Software Quality*, 2022. doi: https://doi.org/10.1007/978-3-030-98464-9_3.

[13] J. Kephart and D. M. Chess, "The Vision of Autonomic Computing," *Computer*, vol. 36, pp. 41–50, 2003, doi: https://doi.org/10.1109/MC.2003.1160055.

[14] P. Arcaini, E. Riccobene, and P. Scandurra, "Modeling and Analyzing MAPE-K Feedback Loops for Self-Adaptation," in *2015 IEEE/ACM 10th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2015, pp. 13–23. doi: https://doi.org/10.1109/SEAMS.2015.10.

[15] "IEEE Standard Glossary of Software Engineering Terminology," *ANSI/ IEEE Std 729-1983*, vol. 0, no. , pp. 1–40, 1983, doi: https://doi.org/10.1109/IEEESTD.1983.7435207.

[16] C. Konstantinou, M. Maniatakos, F. Saqib, S. Hu, J. Plusquellic, and Y. Jin, "Cyber-physical systems: A security perspective," in *2015 20th IEEE European Test Symposium (ETS)*, 2015, pp. 1–8. doi: https://doi.org/10.1109/ETS.2015.7138763.

[17] J. F. Woodward, "Scientific Explanation," *The British Journal for the Philosophy of Science*, vol. 30, pp. 41–67, 1979, doi: https://doi.org/10.1093/bjps%2F30.1.41.

[18] P. Ghazi and M. Glinz, "Challenges of working with artefacts in requirements engineering and software engineering," *Requirements Engineering*, vol. 22, pp. 359–385, 2017, doi: https://doi.org/10.1007/s00766-017-0272-z.

[19] P. Brown, "Reference Model for Service Oriented Architecture 1.0," 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:61416862

[20] E. Mnkandla, "About software engineering frameworks and methodologies," in *AFRICON 2009*, 2009, pp. 1–5. doi: https://doi.org/10.1109/AFRCON.2009.5308117.

[21] L. Chazette and K. Schneider, "Explainability as a non-functional requirement: challenges and recommendations," *Requirements Engineering*, vol. 25, pp. 493–514, 2020, doi: https://doi.org/10.1007/s00766-020-00333-1.

[22] A. Bairy, W. Hagemann, A. Rakow, and M. Schwammberger, "Towards Formal Concepts for Explanation Timing and Justifications," *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, pp. 98–102, 2022, doi: https://doi.org/10.1109/REW56159.2022.00025.

[23] L. Chazette, V. Klös, F. A. Herzog, and K. Schneider, "Requirements on Explanations: A Quality Framework for Explainability," *2022 IEEE 30th International Requirements Engineering Conference (RE)*, pp. 140–152, 2022, doi: https://doi.org/10.1109/RE54965.2022.00019.

[24] M. Camilli, R. Mirandola, and P. Scandurra, "XSA: eXplainable Self-Adaptation," *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 2022, doi: https://doi.org/10.1145/3551349.3559552.

[25] M. Winikoff, G. Sidorenko, V. Dignum, and F. Dignum, "Why bad coffee? Explaining BDI agent behaviour with valuings," *Artif. Intell.*, vol. 300, p. 103554, 2021, doi: https://doi.org/10.1016/J.ARTINT.2021.103554.

[26] I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Modeling and User-Adapted Interaction*, vol. 27, pp. 393–444, 2017, doi: https://doi.org/10.1007/s11257-017-9195-0.

[27] B. K. Kahn, D. M. Strong, and R. Y. Wang, "Information quality benchmarks: product and service performance," *Communications of the ACM*, vol. 45, pp. 184–192, 2002, doi: https://doi.org/10.1145/505248.506007.

[28] M. Nauta *et al.*, "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," *ACM Computing Surveys*, vol. 55, pp. 1–42, 2022, doi: https://doi.org/10.1145/3583558.

[29] H. Balzert, "Lehrbuch der Softwaretechnik - Basiskonzepte und Requirements Engineering, 3. Auflage," in *Lehrbücher der Informatik*, 2009. doi: https://doi.org/10.1007/978-3-8274-2247-7.

[30] H. A. Simon, *The Sciences of the Artificial*. The MIT Press, 2019. doi: https://doi.org/0.7551/mitpress/12107.001.0001.

A. Appendix

A.1. Background

A.1.1. Types of Explananda

Table 12: Types and Examples of Explananda following Chazette et al. [6].

Characteristic	Example
a system in general	-
its reasoning processes	inference processes for problems
its inner logic	relationships of in- and outputs
its model's internals	parameters and data structures
its intention	pursued outcome of actions
its behaviour	real-world actions
its decision	underlying criteria
its performance	predictive accuracy
its knowledge	user / world knowledge

A.1.2. Different Definitions of Explanations

Proposition 2 *(Explanation by [4]): E is an explanation of* *explanandum (X)* *with respect to* *aspect (Y)* *for* *target group (G)*, *in* *context (C)*, *if and only if the processing of E in C by any* *representative (R)* *of G makes R understand X with respect to Y.*

Proposition 3 *(Explanation by [7]): An* *explanation (E)* *for a given* *explanandum (X)* *and a* *target group (G)* *of stakeholders is* ***a piece of*** ***information (I)*** ***(or evidence)*** *that makes the X* ***interpretable*** *by G.*

Proposition 4 *(Explanation by [1]): An explanation (E) for a given explanandum (X) and a target group (G) of stakeholders* ***with an explainability goal (O)*** *is a piece of information (I) (or evidence) that makes the X* ***understandable*** *by G with respect to O.*

Proposition 5 *(Explanation by [10]): An explanation (E) is a piece of information (I) that* ***contributes*** *to the addressee's (A) understanding of an explanandum (X) in a context (C).*

A.1.3. Different Definitions of Explainability

Proposition 6 *(Explainable System by [4]):* ***A system (S) is explainable by means (M)*** *with respect to aspect (Y) of an explanandum (X), for a target group (G) in context (C), if and only* ***if M is able to produce an E*** *in C such that E is an explanation of X with respect to Y, for G in C. [4]*

Proposition 7 *(Explainable System by [6, 10]): A system (S) is explainable with respect to an explanandum (X) of S relative to an addressee (A) in context (C) if and only if there is* ***an entity (T) (the explainer)*** *who, by giving a corpus of information (I) (the explanation (E) of X),* ***enables A to understand X*** *of S in C.*

Proposition 8 *(Explainable System by [7]): A system (S) is explainable if, and only if, it is able by a means (M) to produce an explanation (E) of an explanandum (X) for a target group (G) in a certain operating context (C).*

Proposition 9 *(Explainable System by [1]): A system (S) is explainable if, and only if, it is able to produce by a means (M) an explanation (E) of an explanandum (X) for a target group (G),* ***with an explainability goal (O)*** *in a certain operating context (C).*

A.1.4. Research Fields regarding Explainability

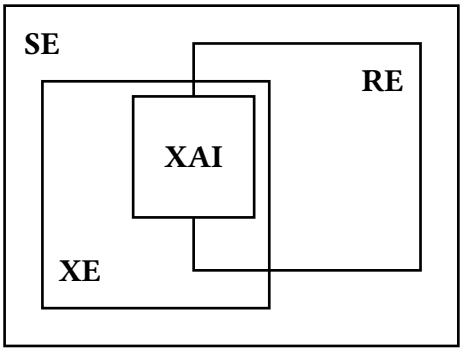

Figure 4: Selected Research Fields in Software-Engineering.

A.2. Related Work

A.2.1. Definitions for Models

Proposition 10 *(Explanation Model (EM)): A behavioural model of the system that captures causal relationships between events and system reactions. [9]*

Proposition 11 *(System Model (SM)): A system model is an artefact from Software Engineering processes, e.g. architecture diagrams, communication protocols, etc. [3]*

Proposition 12 *(Context Model (CM)): A context model, also called environment model Env, describes the operating context of a system. [3]*

A.2.2. Explanation-Model Generation

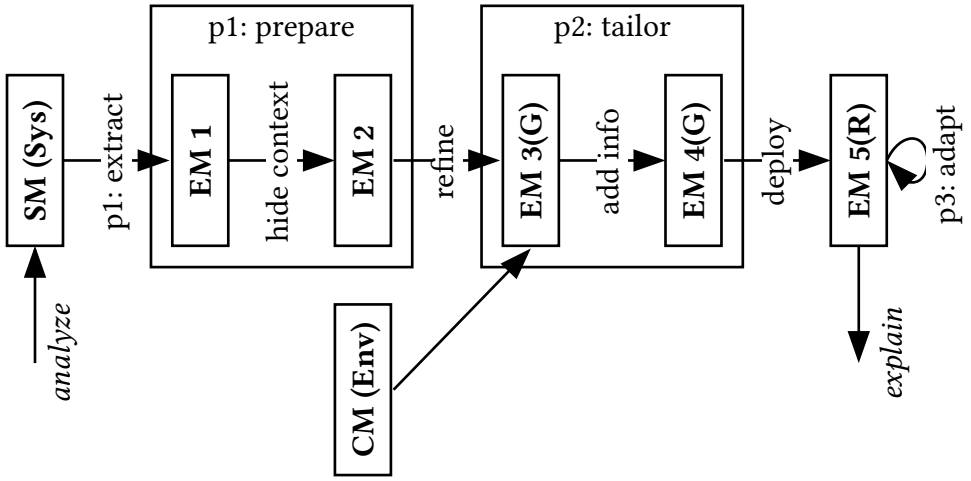

Figure 5: The multi-level extraction and refinement process adapted from Schwammberger et al. [2] and Schwammberger [3]; this could be used in MAB-EX's build phase.

A.2.3. Explanation Justification

Before an explanation was given, a possible phenomenon $p \in P$ to be explained isn't covered by M_A yet (i.e. the phenomenon isn't sufficiently understood), but is by W and should be by M_T in order to be explainable by T. In order for information produced by T and transmitted to A as messages m_i to count as explanation for p, $p \in M_A$ must hold after a set of those messages M has been comprehended by A.

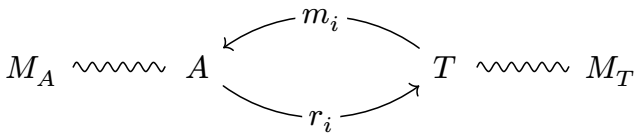

Figure 6: Simple Visualization of Explanations as Interactions; addressee A and explainer T interchange messages m and responses r; both have World-Models M; a phenomenon p is initially: $p \in M_T$ and $p \notin M_A$.

Because explaining is - as already stated - neither a single step [22], nor does it monotonically increase A's *distance to their goal*, T needs to monitor A's reaction to m_i and adapt m_{i+1} accordingly.

A.2.4. Definitions of Explanation Correct- and Goodness

Proposition 13 *(Explanation Correctness by [3]): An internal explanation (E) is a correct explanation E_c, if it can be deduced for an explanandum (X) and from provably correct system models (SMs) and environment models (CM).*

Proposition 14 *(Explanation Goodness by [3]): An explanation (E) can be labelled as good, if it stems from a correct explanation E_c and measurably helps a target group (G) in understanding an explanandum (X).*

A.2.5. Levels of Explainability

Table 13: Levels of Explainability adapted from Bersani et al. [7].

Level	Description	Meta-Requirements
L1	no explainability	-
L2	recognition of needs	A63, M73, M67
L3	local explainability: single/multiple agents	A63, M73, M67 M74, A64 / M66, G61
L4	global explainability: single/multiple agents	A63, M73, M67, M74 A65 / G62

A.2.6. Explanation Quality

Proposition 15 *(Explainability Requirement by [7]): The explanation $E(L_i, X, C, G)$ at a Level L_i, derived for an explanandum (X), in a context (C) for stakeholders (G), shall have quality greater than a threshold $\varepsilon \in \mathbb{R}_+$ established by target group (G): $Q_{E(M)} \geq \varepsilon$.*

Proposition 16 *(Explanation Quality by [1]): The explanation (E), derived for an explanandum (X), with respect to the explainability goal (O), shall have the quality Q(E, X, O). This quality corresponds to the degree with which E aids O of the target group (G).*

A.3. Requirements Analysis

A.3.1. All theoretical Requirements in full Length

Presence of Software System: (P1) (conclusion of X9)
A system with software components must exists.

Presence of Means: (P2) (conclusion of X9, derived from XT)
The system must have means to produce an explanation [1, 4, 7, 10].

Presence of Explainability Goal: (P3) (derived from XT, MAB-EX)
The explain system's content must have an explainability goal. [1]

Presence of Context: (P4) (conclusion of P6, P5, P8, OP7, M66 derived from XT, MAB-EX)
The control system has some context in general. [1, 4, 7, 10]

Presence of Phenomena: (P5) (conclusion of M74, derived from XT and MAB-EX)
The control system shows phenomena in its context. [1, 3, 4, 7, 10]

Presence of Stakeholders: (P6) (conclusion of OP7, P8, OM75)
The control system has different stakeholders.

Presence of Stakeholder Goal: (OP7) A stakeholder should have a goal when using the explain system. [26]

Presence of Explanation Needs: (P8) The control system's phenomena cause different explanation needs.

Producibility of Content: (X9) (derived from XT, MAB-EX)
The explain system must be able to produce information. [4, 9]

Benefit of Content: (G10) (conclusion of G40, derived from XT)
The information produced should contribute to a stakeholder understanding explananda of the control system. [3, 27]

Understandability of Content: (G11) (conclusion of G10, G12, G13)
The information produced must be understandable by a stakeholder.

Comprehensibility of Content: (G12) (derived from XT)
The information's presentation should be understandable by a stakeholder. [27]

Grasping of Content: (G13)The information's meaning should be understandable by a stakeholder.

Adaptability of Information: (G14) The information should be adaptable during production.

Preferences of Stakeholder: (B15) (conclusion of X16, X17, derived from XT)
The explain system should be able to build appropriate displays of information for stakeholders. [23, 26, 27]

Adaptability of Presentation Syntax: (X16) (conclusion of G12, OX20, OX19, OX20, derived from XT, MAB-EX)
The syntax of information presentation must be adaptable. [9, 10, 23, 26–28]

Adaptability of Presentation Semantics: (X17) (conclusion of G13, OX23, OX24, OX22, OX21, derived from XT)
The semantics of information presentation must be adaptable. [23, 26–28]

Adaptability of Media: (OX18) (derived from XT)
The medium, in which information is expressed, should be adaptable. [23, 26]

Adaptability in Media-Amount: (OX19) (derived from XT)
The amount of different media types should be adaptable. [23]

Adaptability in Amount: (OX20) (derived from XT)
The *size* (or length) of information should be adaptable. [27, 28]

Adaptability of Relevance: (OX21) (derived from XT)
The information should be of relevance to the phenomena of interest. [27, 28]

Adaptability of Contrast: (OX22) (derived from XT)
The information should be adaptable in its contrast, i.e. the scope of information included in it. [23, 28]

Adaptability of Complexity: (OX23) (derived from XT)
The information should be adaptable in its complexity. [27, 28]

Adaptability of Style: (OX24) (derived from XT)
The communication of information should be adaptable in its style. [23, 26]

Buildability of Explanation-Model: (B25) (conclusion of G36, B43, E48, B50, B47, derived from XT, MAB-EX)
The explain system must be able to build a model as a basis for information production. [1, 2, 9]

EM Initializability: (B26) (derived from XT)
An explanation model should be extractable from a system model. [2]

Manual Constructability: (OE27) (conclusion of X9, derived from XT, MAB-EX)
The explanation model should be manually constructable. [2, 9]

Automated Constructability: (OE28) (conclusion of X9, derived from XT, MAB-EX)
The explanation model should be automatically constructable. [2, 9]

Complete Constructability: (OE29) (conclusion of G39, derived from XT)
The system model should include every possible system behaviour. [2]

Correct Constructability: (OE30) (conclusion of G38, derived from XT)
The system model should include system behaviour accurately and correctly. [2]

EM Relevance: (E31) (derived from XT)
Details not relevant for the explainability goal should be hideable. [2]

EM Generalizability: (E32) (derived from XT)
The explanation model should be generalizable towards a general target group. [2]

EM Individuality: (E33) (conclusion of B15, derived from XT)
The explanation model should be individualizable towards a specific representative of a target group. [2]

EM Contextualization: (E34) (derived from XT)
The explanation model should be extensible by stakeholder-specific information from a context model. [2]

Information Coherence: (G35) (derived from XT)
The information produced should be based on the stakeholder's knowledge base. [28]

Information Objectivity: (G36) (derived from XT)
The information produced shall be unbiased, unprejudiced and impartial. [27]

Information Security: (E37) (derived from XT)
The explanation model should be adaptable based on a stakeholder's privileges to ensure data security. [27]

Information Correctness: (G38) (conclusion of G40, derived from XT)
Information generated shall be deduced from provably correct system models and context models. [3, 27, 28]

Information Completeness: (G39) (derived from XT)
Information generated needs sufficient breadth, depth and amount to fulfil explainability goals. [27, 28]

Information Goodness: (G40) (derived from XT)
Information must be correct and measurably help a target group to understand an explanandum. [3]

Information Consistency: (G41) (conclusion of G38, derived from XT)
The generation of information must be the same under the same conditions [27, 28], i.e. be deterministic.

Information Continuity: (OG42) (conclusion of G38, derived from XT)
The generation of information should be similar, if the conditions change only slightly [28], i.e. be continuous.

External Triggerability: (B43) (conclusion of G51, derived from MAB-EX)
The production of information by the explain system must be externally triggerable (e.g. by stakeholders) [9].

Internal Triggerability: (B44) (conclusion of G51, derived from MAB-EX)
The production of information by the explain system must be internally triggerable (e.g. by an analysis) [9].

Detectability of Explanation Needs: (A45) (conclusion of E33, E32, B43, B44, derived from XT, MAB-EX)
The explain system should be able to detect explanation needs. [7, 9]

Extractability from Explanation Model: (B46) (derived from XT, MAB-EX)
Information in the explanation model must be extractable. [2, 9]

Simulation of Behaviour: (B47) (conclusion of G51, derived from XT, MAB-EX)
Future explananda of the control system should be simulatable based on the explanation model. [9, 22]

History of Explanation Model: (E48) (derived from MAB-EX)
The explain system should keep a chronological history of past explanation models. [1]

Information Recency: (B49) (conclusion of G10, derived from XT)
The information produced by the explanation model should be as recent (up-to-date) as possible. [27]

Updatability of Explanation Model: (B50) (conclusion of G51, X16, X17, derived from XT)
The explanation model adapted for a specific stakeholder should be updatable at run-time. [2]

Information Timing: (G51) (conclusion of G10, derived from XT)
An information should be adaptable in its timing, i.e. the occurrence of it before, during, or after an event [22].

Parallel Production: (OB52) (conclusion of G51)
Information for several stakeholders should be producible in parallel.

Prioritizability in Production: (OB53) (conclusion of G51)
The production of information for a stakeholder should be prioritized based on the urgency of the need.

Computability of Benefit Maxima: (A54) (conclusion of G51, derived from XT)
The information-timing with the highest benefit for a stakeholder should be computable. [22]

Computability of Need-Importance: (OA55) (conclusion of OM75, A45, derived from XT)
The importance of an explanation to a stakeholders need shall be computable [22].

Analytic Capability: (A56) (conclusion of B50)
The explain system should be capable of analysing data it observes.

Analysis of Goals: (A57) (derived from MAB-EX)
The goals of stakeholders should be retrievable by analysing the interactions with, and behaviour of the system [9].

Analysis of Understanding: (A58) (conclusion of MAB-EX)
A stakeholder's feedback should be analysable in order to verify if an information was understood. [9]

Computability of Need: (A59) (conclusion of A45)
A stakeholder's explanation need should be computable.

Analysis of Interactions: (OA60) (conclusion of B50, derived from MAB-EX)
Interactions of the system and stake-holders should be analysable for aimlessness or contradictions [9].

Local Explanation: (G61) (derived from XT)
A local explanation for an individual explanandum shall be producible by considering all partial contexts. [7]

Global Explanation: (G62) (derived from XT)
A global explanation of the average of an ex-planandum shall be producible by considering all partial contexts. [7]

Factorization of Context: (A63) (conclusion of G61, G62, derived from XT, MAB-EX)
The context shall be analysable as factors affecting the explanandum in the system's explainability goal. [7, 9]

Computability of Individual: (A64) (conclusion of G61, G62, derived from XT)
An individual explanandum shall be computable as the sum of effects of the observable partial contexts. [7]

Computability of Average: (A65) (conclusion of G61, G62, derived from XT)
The average of an explanandum shall be computable as the expected distribution of it. [7]

Partial Monitorability: (M66) (conclusion of A63, derived from XT)
Each partial context shall be monitorable by the corresponding agent participating. [7]

Behavioural Measurement-Availability: (M67) (conclusion of A63, A45, derived from XT)
A measurement of the context-factors shall be available when the explananda occur. [7]

Presence of Interfaces: (P68) (conclusion of M70, M69)
The control system should have interfaces with the context.

Capturability of Interface-Events: (M69) (conclusion of A56)
Possible interfaces of the control system must be observable by the explain system.

Capturability of Sensor Data: (M70) (conclusion of A56, derived from MAB-EX)
The explain system should be able to capture observer data used by the control system. [9]

Capturability of Commands: (M71) (conclusion of A56, derived from MAB-EX)
The explain system must be able to capture (former) commands used by the control system. [9]

Capturability of Explanations: (OM72) (conclusion of A56, E48, derived from MAB-EX)
The explain system should be able to capture (former) explanations. [1]

Monitorability of Context: (M73) (conclusion of A63, derived from XT, MAB-EX)
The context shall be observable and measurable at runtime. [7, 9]

Monitorability of Explananda: (M74) (conclusion of A56, derived from XT, MAB-EX)
The control system's explananda must be observable [7, 9] for stakeholders, and explain systems [9].

Monitorability of Stakeholder: (OM75) (conclusion of A45, derived from MAB-EX)
The explain system should be able to monitor a stakeholder including its behaviour and feedback. [9]

Monitorability of Interactions: (M76) (conclusion of OA60, derived from MAB-EX)
The explain system must be able to monitor interactions of the control system and stakeholders. [9]

Chronological Capturability: (OM77) (conclusion of A56, E48)
The explain system should be able to capture the chronical aspects of all captured events.

A.4. Case-Study

A.4.1. Requirements Elicitation Practices

Table 14: Practises defined by Chazette [8].

1. Definition of Visions (and Goals) → Section 6.1
2. Trade-Off-Analysis → *not in scope*
3. Stakeholder-Analysis → Section 6.3
4. Backend-Analysis → *not in scope*
5. Explainability-Design → *not in scope*
6. Evaluation → *not in scope*

A.4.2. All practical Requirements in full Length

Deviation Detectability (C1) (based on S-1) The system must be able to detect deviations in its context. Therefore, it must have the ability to observe and capture the context, as well as some understanding or representation of what counts as a deviation and what is normal. → confirms: P4, A63, M73

Timing of Explanations (C2) (based on S-1) In order to maximize the benefit of an explanation for an addressee, the system must be able to give the information at the right time. For this, it must also be able to compute a point in time with maximized benefit, and to produce explanations in parallel. → confirms: G51, OB52, A54, M67

Level of Detail (C3) (based on S-2) The system should be able to detect what level of detail is needed in a potential information for a target group. → confirms: OX23

Determinism (NFC4) (based on S-3) The generation of information by the system should be consistent and continuous, i.e. if the context, including the target group, is the same, the resulting information must also be the same. Additionally, they should be similar, if the inputs were similar. → confirms: G41, OG42

Contrast (C5) (based on S-4, S-7) In order to let a representative of a target group understand the reasons to *why* an explanandum (X) happened, information on the contrast of X, e.g. *why not* another X happened, should be included. For this, *varying contexts* should be considered. → confirms: OX22, G61, G62, A65, M66, M74

Simulatability of Behaviour (C6) (based on S-1, S-5) In order to provide information about the consequences of deviations, the explain system should be able to simulate the future behaviour of the context. Therefore it needs a time or state model. → confirms: B25, B46 , B47

Prioritizability of Generation (C7) (based on S-6) When interacting with multiple target groups, the system must analyze their priority to ensure that those with the greatest need receive their necessary information first. → confirms: OB53

Urgency of Information-Need (C8) (based on S-6) In order to maximize the benefit of the information generated, the system must be able to compute the urgency of the target group's need. → confirms: OA55

Representation of Content (C9) (based on S-6, S-7) The information generated by the system must be adaptable in its presentation. Hence, the system must be able to identify the most appropriate way to present information to a target group. → confirms: X16, OX18, OX19, OX20, OX24

Triggerability of Generation (C10) (based on S-7) The generation of information by the system must be triggerable externally through the explicit expression of a target group or internally when an implicit explanation need is detected. Thus it must also have an interface where a need can be detected. → confirms: B43, B44, A45, A59, P68, M69

Correlations of Events (C11) (based on S-8, S-9, S-10, S-11, S-12) In order to generate explanation for certain behaviour, the system must be able to detect and capture correlations between events in the context and resulting ones of the control system. → confirms: B25, B26, E34

History of Behaviour (C12) (based on S-9, S-10, S-11, S-12) In order to generate an explanation of the control system's past decisions and behaviour, it must be able to capture and store relevant data (logs, models, explanations) in the correct historical order. → confirms: E48, M70, M71, OM72, OM77

Data Privacy/Security (NFC13) (based on S-10, S-11, S-12) The sensitive data collected and stored by the explain system must always be handled under strict privacy and security measures in order to ensure stakeholder trust in the system. → confirms: E37

Validity of Information (NFC14) (based on S-6) The information generated by the system must be provably valid in order to be beneficial for the target group and to build trust in the system. Thus, also the information used as a basis for build models must also be valid. → confirms: OE29, OE30, G38, G39

Explainability (NFC15) (based on *all cases*) The system must be explainable for a target group in a context with respect to an aspect of an explanandum [4]. → confirms: P5, M74, OM75

Producibility of Information (C16) (based on *all cases*) The system should be able to produce some information in order to address potential needs of target groups. Therefore it has to have some means. → confirms: P2, X9

Real-Time Generation (C17) (based on *all cases*) The information required by target groups must be generated at run-time by the system in order to be up-to-date. For this reason, also prior system interactions with target groups must be taken into account in potential future information-generation. → confirms: B49

Individuality of Information (C18) (based on *all cases*) To maximize the benefit of an explanation for an addressee (A), the system should adapt the information to align with A's current (world) knowledge. The information should be coherent with A's knowledge, but should not contain any bias towards A, and therefore be objective. → confirms: G14, B15, G35, G36, A56, A64

Benefit of Information (C19) (based on *all cases*) Through receiving an explanation, the distance to a threshold representing the understanding of an explanandum shall decrease in order for the explanation to be beneficial to the addressee's understanding of the explanandum. → confirms: G10

Measurability of Understanding (C20) (based on *all cases*) In order to measure whether the needs of a target group were satisficed or whether further explanations are needed, the system should be able to evaluate whether the target group has understood the information provided. → confirms: G11, G12, G13, A58

Relevance of Information (C21) (based on *all cases*) The information produced by the system should be of relevance in order to be beneficial to the goals of both entities. Thus, the system must be able to analyze the addressee's goals or motivations and (former) interactions between them. → confirms: OX21, A57, OA60, M76

A.5. Evaluation

A.5.1. Remarks of the Evaluation

Table 15: All Remarks derived in the Comparison.

Remark 1: All presence requirements are either confirmed directly or implied indirectly.
Remark 2: All general requirements are confirmed by case-requirements. (Regardless of whether they are explicitly required by MAB-EX)
Remark 3: All explanation requirements are confirmed by case-requirements.
Remark 4: Details on information-presentation could be used to refine MAB-EX.
Remark 5: Some technical build requirements are not confirmed by case-requirements.
Remark 6: All build requirements required by MAB-EX are confirmed by case-requirements.
Remark 7: Some technical explanation model requirements are not confirmed by case-requirements.
Remark 8: All monitor- and analysis-requirements required are confirmed by user needs.
Remark 9: Each case-requirement is covered by at least one theory requirement.
Remark 10: Not all case-requirements are covered by theory requirements that are explicitly stated in a paper regarding the development of MAB-EX.
Remark 11: It might be beneficial to incorporate the proposed insights on explanation-timing, -justification, -correctness and -models into MAB-EX.
Remark 12: Aspects connected to data security and the priority in production might be needed in MAB-EX.
Remark 13: The idea of an explanation quality incorporated into MAB-EX is beneficial for self-explainability, supporting the understanding by stakeholders.