Towards the Operationalization of Trustworthy AI: Integrating the EU Assessment List into a Procedure Model for the Development and Operation of AI-Systems

Henrik Kortum[1], Jonas Rebstadt[1], Tula Böschen[1], Pascal Meier[1], Oliver Thomas[1]

Abstract: Artificial intelligence (AI) is increasingly permeating all areas of life and not only changing coexistence in society for the better. Unfortunately, there is an increasing number of examples where AI systems show problematic behavior, such as discrimination or insufficient accuracy, missing data privacy or transparency. To counteract this trend, an EU initiative has drafted a legal framework and recommendations on how AI can be more trustworthy and comply with people's fundamental rights. However, fundamental rights are currently not reflected in procedure models for the development and operation of AI systems. Our work contributes to closing this gap so that companies, especially SMEs with small IT departments and limited financial resources, are supported in the development process. Within the framework of a structured literature review, we derive a procedure model for the development and operation of AI systems and subsequently integrate concrete recommendations for achieving trustworthiness.

Keywords: Trustworthy AI, Procedure model, Explainable AI, Machine Learning, SME

1 Motivation

Artificial intelligence (AI) and Machine Learning (ML) are permeating all areas of life, changing the way we live together in society, not only for the better [WaSi18]. There is an ever-growing number of examples in which AI systems have shown behavior that is inconsistent with fundamental rights [CaCS20, VaWi18]. For instance, they reinforce existing discriminatory biases [RoBL20], cannot meet high quality requirements in practical use [LeLL21] and omit interests of useres with respect to data privacy, transparency and autonomy. Examples include AI systems in recruiting that discriminate against females, certain ethnic groups or people with disabilities [BaHN17] or chatbots with racist, sexist, and antisemitic tendencies [WoMG17]. Moreover, the extended application of ML establishes new possibilities for cybercrime, e.g. by injecting adversarial examples using perturbations of the input vector which are in some cases not even noticeable by humans and lead to undesired behavior of the AI system [YHZL19]. In order to address those deplorable conditions, the European Union (EU) has specified regulatory requirements and developed a legal framework for the regulation of AI systems [Euro21] that is expected to have a massive impact on AI deploying and developing companies. Up until now, research in the field of ethical AI has been predominantly

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH – Smart Enterprise Engineering, Parkstraße 40, 49080 Osnabrück, {henrik.kortum, jonas.rebstadt, tula.boeschen, pascal.meier, oliver.thomas}@dfki.de

doi:10.18420/inf2022_26

conceptual and focused on defining principles [SeBM21]. But as stated by [MiCo19] 79% of all tech workers demand specific, practical resources regarding ethical considerations. Small and medium-sized enterprises (SMEs) in particular, where the use of AI is far less widespread than in larger companies [BaDK20], will face further challenges from increasing regulation of AI. However, such regulation also offers competitive opportunities, especially for European companies [FeDe21] and SME [TiOH21]. Following this argumentation, the goal of this paper is to increase the potential for simple, guided ethical consideration by mapping the ALTAI [KoGe20] assessment questions onto a practical procedural model. In current practice, the development of AI systems is often based on the use of established procedure models, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) or Knowledge Discovery in Database (KDD). However, these models do not take into account how AI systems should be designed in a trustworthy way to specifically avoid discrimination. Our work attempts to resolve this research gap by answering the following research question:

RQ: How can a procedure model for the development and operation of trustworthy AI systems be designed in accordance with the Assessment List for Trustworthy AI (ALTAI)?

Within a structured literature review, we identify relevant phases and activities from existing procedure models for the development and operation of AI systems. From the results of the literature review we subsequently derive an integrated procedure model. The resulting model is supplemented with concrete guidelines for the development of trustworthy AI systems, which we derived from the ALTAI. In doing so, we add value to both the IS research community and practitioners, providing them with the procedural model for developing and operating trustworthy AI systems. The responsible expert group of the European Commission proposes the establishment of cross-functional teams, consisting of AI developers, data scientists, procurement officers, front-end staff, legal/compliance officers and managers, for the implementation of the ALTAI. For smaller companies, however, such project stuffing hardly seems feasible, which is why we would like to explicitly address these companies with our work. The rest of our paper is structured as follows. Section 2 provides an overview of relevant literature. In Section 3 we introduce our design science-based research approach. In 4.1, we derive an integrated procedure model based on the findings of a structured literature review Section 4.2 presents the core aspects of ALTAI and provides the foundation of our requirements for trustworthy AI systems development. Subsequently, in Section 4.3, we combine the model from 4.1 with the requirements from 4.2 and propose our final model. We discuss the results in Section 5 and finalize with possible future directions in Section 6.

2 Procedure Models and Trustworthiness of AI Systems

The development of software and AI systems can be structured in typical activities and phases as is done by procedure models in classical software development. Likewise to the waterfall model [Royc87] in classical software development, there are some established

procedure models for data mining and the development of AI systems. For example, KDD (Knowledge Discovery in Databases) was one of the first models addressing the particularities of data mining [FaPS96, MaMF10]. [CCKK00] introduced CRISP-DM, which is widely used in practice It provides a standard for data mining projects and comes with relevant implications for the development of AI systems. In CRISP-DM, the essential phases "Business Understanding", "Data Understanding", "Data Preparation", "Modeling", "Evaluation" and "Deployment" are differentiated, including constant feedback between the different phases. Besides KDD as foundation for most current models and CRISP-DM already merging phases from multiple models, there is a variety of models with a holistic or a domain specific approach like [SiSS19] for the public mental health sector [MaMF10]. Alongside the demand for efficient development, the representation of ethical and regulatory implications becomes increasingly important in procedure models, especially since the EU's publication of a draft for a legal framework [Euro21].

The EU embosses the term of ethical or trustworthy AI. It summarizes a selection of concepts and principles, which constitute the robust, legally secure, and ethically defensible development and use of AI systems. Even though the EU's Guideline for Trustworthy AI initially defines its principles in an abstract manner as well, it is one of the first guidelines that offers an intuitive resource for practioneers, namely ALTAI. On an abstract level, the guideline is based on three key concepts: Lawfulness, ethics, and robustness [High18]. The ethical perspective is broken down into four principles. (1) *Respect for human autonomy* ensures that AI systems enhance, complement, and promote human capabilities without limiting their freedom or autonomy. (2) *Prevention of harm* focuses on the mental and physical integrity of humans interacting directly or indirectly with AI systems. (3) *Fairness* represents the focus of this publication and puts forward fairness and non-discrimination in the development, deployment, and use of AI systems. The aim here is to ensure equal opportunities on the one hand, but also to create the possibility of appealing against decisions made by the systems. The starting point for this is the transparency of the AI system's decision-making process, which is made explicit by the principle of (4) *explicability*. According to this principle, the purpose and capabilities of the AI system should be disclosed, and it should be possible to explain the decisions to all affected persons. The practical specification of these principles, including the resulting questions relevant for the stakeholders involved in the development process of AI-systems (ALTAI), comes with high ethical integrity for the following reasons: ALTAI was published by the Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) which was set up by the European Commission. The EU submits to basic ethical values that are represented as fundamental rights. The expert group states that its motivation to encourage trustworthy AI is the compliance of fundamental rights. Members of AI HLEG come from diverse backgrounds, so that economy and science, legal authorities and different nationalities are represented in their views and interests. Also, an earlier version of the document went through a multimodal piloting phase such that broad feedback is incorporated to the final version of the assessment list. Our goal is to increase the potential for simple, guided ethical consideration by providing a mapping of ALTAI's assessment questions to a procedure model. We employ ALTAI as a basis because it

comes in a form that matches our line of thought and progress: Since ALTAI is designed as a list for self-assessment, it is framed in a user-friendly, concrete way. Additionally, its specific relevance for the European community and its high level of congruence to the principles mentioned by [JoIV19], encourage our choice. Although ALTAI is of integrity with respect to its development and design, it has to be pointed out that it does not claim to be complete. Our literature review supports that different ethical assessment lists are hardly congruent. Their focus depends on the field of application as well as the perspectives of its authors and ethical priorities that can also be contingent on time, culture and more. A selection process of headwords necessarily requires some ethical concerns to be disregarded. Also, during the application of an assessment list, trade-offs between the principles might present themselves as unavoidable. These kinds of choices reflect embeddedness of authors and developers in society, culture and other circumstance and should be continuously and critically monitored.

3 Research Approach

In order to derive an integrated procedure model for the development and operation of AI systems, a structured literature search according to [WeWa02] was carried out. Aim of the conducted literature search was the identification of relevant phases and activities of procedure models for the development and operation of AI systems. For this purpose, the following search term was applied to eight relevant databases in the field of information systems research: (*"procedure model" OR "process model" OR lifecycle OR framework) AND ("data science" OR "data mining" OR "data engineering" OR "machine learning" OR "artificial intelligence" OR "knowledge discovery" OR "decision support systems")*.The results were analyzed, based on the relevance of the publications in relation to our research question, by first analyzing the title, then the abstract and finally the full text. Tab. 1 shows the hit quantities of the individual analysis steps per database. The remaining 80 relevant publications were analyzed and categorized based on [WeWa02], identifying 7 phases and 44 relevant activities as presented in Section 4.1.

Database	Hits	On title	On abstract
IEEE	239.760	118	42
AISel	18.257	27	17
SpringerLink	176.995	82	29
Science Direct	133.311	46	21
Wiley	171.846	56	13
Ebsco	2	0	0
Web of Science	89.509	71	32
JSTOR	0	0	0
		Summed up	**154**
		Duplicates, no access or out of scope	**74**
		Total	**80**

Tab. 1: Overview: Literature review

In order to determine the trustworthiness requirements of AI services, we analyzed the ALTAI in detail (Section 4.2). The subsequent mapping of the integrated process model and ALTAI was carried out independently by three experts and consolidated within a workshop. The method of theoretical reasoning was used for this purpose. Outcome of the workshop and answer to our research question is our proposed procedure model for the development and operation of trustworthy AI systems (Section 4.3).

4 Derivation of a Procedure Model for the Development and Operation of Trustworthy AI Systems

4.1 Integration of existing Procedure Models from the Literature

As described in Section 3, we conducted an extensive search of existing models as part of a structured literature review to identify relevant phases and activities. Over the years several procedure models for AI system development have been developed. Even though the models differ in several aspects, most of them have an overlap in their core activities. It is worth noting that most of the identified models reference CRISP-DM and are guided by its core structure, which is why we also choose it as the structural starting point of our work. We divided the identified activities into seven main phases based on CRISP-DM, added additional activities and a complementary operations phase to adequately address the recommended actions derived from ALTAI later in 4.3.
The development procedure is initiated by the *Business Understanding* Phase, summing up seven activities. It begins with the activity *Explore Goal* [MCFH19], focusing on the exploration of the reason and the target of the recently initiated development project. Setting up on the defined goal, the background of the problem needs to be evaluated in the activity *Determine Business Objectives* [CCKK00], quantifying the success or the failure of the system, allowing the system to optimize towards the defined goal. As foundation for the project plan, in the *Assess Situation* [CCKK00] activity, resources, costs and benefits are contemplated. To consider possible risks, the activity *Identify Risk Factors* [SiSS19] has been defined, followed by the three final requirement regarded activities: Initially, one needs to *Elicitate Requirements* [ANIN17], followed by *the Requirement Negotiation* [ANIN17] to fix the requirements in a unanimous state and finalized by the *Requirement Validation* [ANIN17]. Building up on the validated requirements, the second phase is *Data Understanding*, the first of two explicitly data-related phases. As initial activity, called *Acquire Data* [MCFH19], all available data is collected. This data needs to be explored in the *Explore Data* activity [MCFH19], described in the *Describe Data* [CCKK00] activity and quality-checked in the activity *Verify Data Quality* [CCKK00]. After the technical evaluation of the data the goal-specific value of the data is quantified in the *Data Value Exploration* [MCFH19]. As foundation for the *Data Preparation* valuable stories are extracted from the data in the *Narrative Exploration* [MCFH19]. In the second explicitly data related phase, *Data Preparation*, the data is set up for the training procedure. First, the relevant data is selected (*Select Relevant Data* [LaRa11]),

cleaned (*Clean Data* [CCKK00]) and variables can be engineered (*Construct Data* [CCKK00]). In case of lacking or unbalanced training data, new data should be generated [MCFH19]. According to task-specific requirements, the resulting data is formalized in the *Formalization of Data* activity and a data architecture is defined (*Define Data Architecture* [MCFH19]), to physically and logically structure the data sources. As final steps, the data is integrated in a single structure (*Integrate Data* [CCKK00]) and formatted into the required structure in the *Format Data* [CCKK00] activity. In the following *Modeling* Phase, the model is developed, starting with the first activity, called *Select Modeling Techniques* [CCKK00] and guarded by the development of a corresponding test design (*Generate Test Design* [CCKK00]). Building up on this, the actual model is developed in the activity *Build Model* [CCKK00]. The model should be compared to an own, task-specific baseline model (*Comparing to Baseline Model* [Wang08]) and assessed in general (*Assess Model* [CCKK00]). As final activity of the current phase, explanatory methods are added, if the model itself is not already inherently interpretable as can be the case in (learned) decision trees or rule-based systems [SaGr17]. To assess and verify the quality of the developed model, the *Evaluation* phase is carried out. The evaluation is done in three main activities, starting with an exploration (*Explore Results* [MCFH19]) and an evaluation (*Evaluate Results* [CCKK00]) of the model-results to address all technical concerns. Next to this, the conducted process needs to be reviewed as well in an activity called *Review Process* [CCKK00]. As one of the most extensive phases in the development of productive AI systems, the *Deployment* contains all activities from the evaluated model towards a releasable system. As initial activity, setting up on the evaluation results, in the *Explore Product Opportunities* activity [MCFH19], the added value and existing possibilities of the developed model in a product setting are assessed. If the development of a product is promising, the activity *Plan Deployment* [CCKK00] is started, followed by the release of the data in the future production environment (*Release Data* [MCFH19]). Using this and the predefined model, the system is implemented (*Implement System* [HSMK19]), incorporating the development of a User Interface (*Implement User Interface (UI)* [LaRa11]) and the inclusion of the developed explanatory approaches (*Integrate Explanations in UI* [SaGr17]). The resulting system needs to be tested as a whole, extending the model-specific evaluation (*Test System* [HSMK19]). After the test procedure, a final report is written (*Produce final Report* [CCKK00]), the system is released (*Release System* [HSMK19]) and finally, the project needs to be reviewed (*Review Project* [CCKK00]). As extension to the main phases of CRISP-DM, the Operation phase has been added in this model, to continuously supervise and optimize the running AI system and incorporate all user related communicatory topics. As initial only activity, *Monitor System* [HSMK19] is focusing on the continuous behavior of the developed system not only once after the development but rather during the whole life cycle of the system. In the following section our proposed model is supplemented by concrete recommendations for action resulting from ALTAI.

4.2 Derivation of a Procedure Model for the Development and Operation of Trustworthy AI Systems

Based on the considerations outlined in Section 2, we decided to use ALTAI as the basis for the assessment regarding trustworthiness, which we include in the procedure model. ALTAI addresses AI developers, project managers, front-end staff and more professionals that work with AI systems. The assessment list has seven sections each of which present a number of headwords. Every headword comes with a set of questions for self-assessment. The seven sections and headwords are presented in Tab. 2. The headwords are named in the brackets and referenced by an identifier. In Section 4.3 we will assign the individual headwords of ALTAI to the appropriate phases and activities of the procedure model in which they are relevant and briefly explain how they can be addressed.

ID	Section Name	Summary and Headwords
R1	Human Agency and Oversight	▪ Assess possible influences of the AI system to individual humans, particularly as the systems guides, influences or supports human decision making (Human Agency and Autonomy, #1.1) ▪ Enable humans to intervene AI system at all times (Human Oversight, #1.2)
R2	Technical Robustness and Safety	▪ Protect the system from physical and cyber-attacks and assess the risks that arise in case of abuse/deficiency (Resilience to Attack and Security, #2.1) ▪ Assess risks that might arise from sloppy design (General Safety, #2.2) ▪ Assess the effects that inaccurate predictions of the system would put forward (Accuracy, #2.3) ▪ Put forward means to compensate for the system in case of failure and ongoingly validate it (Reliability, fallback plans and reproducibility, #2.4)
R3	Privacy and Data Governance	▪ Handle personal (user) data responsibly (Privacy, #3.1) ▪ Assure integrity of data quality and content (Data Governance, #3.2)
R4	Transparency	▪ Assure that the principle of operation and the decisions of the AI system remain traceable (Traceability, #4.1) ▪ Encourage the user's understanding of the AI system's decisions (Explainability, #4.2) ▪ Communicate possible risks and limitations of the AI system to users and, if applicable, provide disclaimers (Communication, #4.3)
R5	Diversity, Non-Discrimination and Fairness	▪ Design data sets and algorithms such that results are fair with respect to diversity and representativeness (Avoidance of unfair bias, #5.1) ▪ Make sure that the system can be used by everyone, including people with special needs or preferences (Accessibility and Universal Design, #5.2) ▪ Consult stakeholders during the development of the AI system (Stakeholder Participation, #5.3)
R6	Societal and Environmental Well-Being	▪ Monitor and reduce negative impacts on the environment (Environmental Well-Being, #6.1) ▪ Monitor impact on the working environment and required skills and make efforts to adapt (Impact on Work and Skills, #6.2) ▪ Monitor and reduce negative impact that the AI system may have on society and democracy (Impact on Society at large or Democracy, (#6.3)
R7	Accountability	▪ Make sure that the system can be audited independently from its development (Auditability, #7.1) ▪ Constantly monitor possible risks that arise in the scope of the AI system and explicate trade-offs between the ethical principles (Risk Management, #7.2)

Tab. 2: Summary of ALTAI [KoGe20]

4.3 Extension of the Procedure Model to include Trustworthiness

In the next step, the aggregated model (see Section 4.1) is enriched by the ALTAI recommendations for action (see Section 4.2), which are assigned to the appropriate phases and activities. It is worth noting that not all requirements derived from ALTAI could be assigned reasonably to any of the activities identified in the literature, which is why we added three new activities to the model. Fig. 1 shows our final procedure model for the development and operation of trustworthy AI systems, with the individual recommendations for actions resulting from the ALTAI grouped by phase described in detail below.

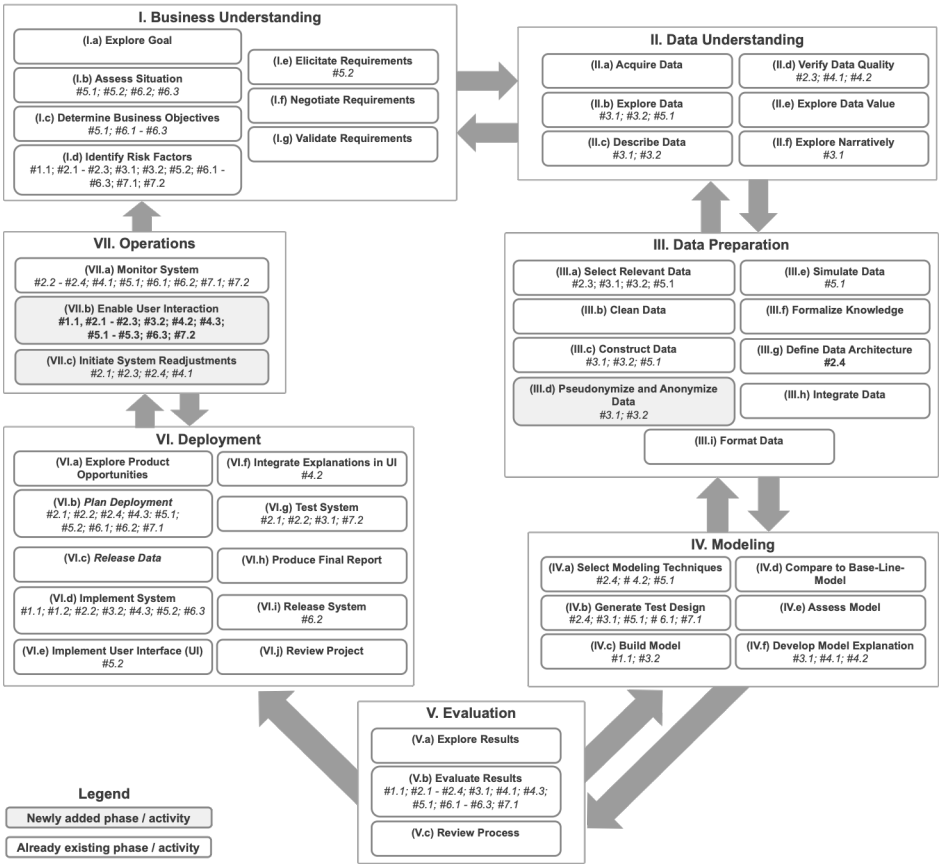

Fig. 1: Procedure model for the development and operation of trustworthy AI-systems

(I) Business understanding: In this phase, requirements of all categories must be considered. When assessing the situation, specific attention should be paid to identify potential biases and to understand the diversity and representativeness of the end-users (#5.1). Even beyond the end users, other affected stakeholders should also be identified

and involved in this activity to be aware of potential discrimination and to include the needs of all stakeholders (#5.2). Of particular importance here are also the workers interacting with the system. AI changes the demands on the labor market [KoRT22], therefore, the potential impact on the skillsets required for the tasks and on work arrangements must be determined (#6.2). It should also be evaluated whether the AI system could have a potential impact on society (#6.3). The next relevant activity, the identification of risk factors, is of central importance for the development of trustworthy AI systems. The relevant risk factors are distributed across all categories of the ALTAI. They include the assessment of possible influences of the AI system on individual humans, especially when the system is guiding human decision making (#1.1). Furthermore, an assessment of the impact caused by misuse and attacks on the AI system (#1.1), sloppy design (#2.2) and inaccurate predictions (#2.3) of the AI system must be made here. Another potential source of risk is inadequate privacy (#3.1) and data protection (#3.2) measures. The risk and consequences of possible inequitable behavior of the system toward end users or individual subgroups must also be assessed (#5.2). In addition to the concrete risks with direct effects, long-term risks with potentially negative consequences for the environment (e.g., due to non-environmentally sustainable operation) (#6.1) and for the working environment (#6.2) of users, and more abstractly, for society as a whole (#6.3) need to be addressed. All these aspects should be ascertainable in a structured way by means of an audit mechanism (#7.1). For organizational guidance, ALTAI recommends third-party consulting, establishing an AI ethics review committee, and providing legal and risk trainings (#7.2). Based on the risk factors, concrete requirements should be identified, negotiated and evaluated. Here, the aspect of accessibility and universal design (#5.2) is decisive. The requirements should correspond to the diversity of preferences and capabilities in society. Special care should be taken to ensure that accessibility to the AI system for persons with disabilities, is adequately addressed in the requirements.

(II) Data understanding: During the exploration of the data, care should be taken to ensure that the data is appropriate and representative of the diversity of the population. In particular, attention should be paid to any bias that may exist that could result in discrimination against certain subgroups (#5.1). Building understanding about the data should be supported by publicly available technical tools. In addition, privacy (#3.1) and data governance (#3.2) aspects must also be considered during this activity. It is necessary to identify data that fall under the Data Protection Regulation (GDPR), or a non-European equivalent, or whose further processing could be critical for other reasons. Therefore, ALTAI recommends that a Data Protection Impact Assessment (DPIA) is carried out here. To allow the classification of critical data to be considered in the further process, the data should be described accordingly. The data quality assessment must ensure that data used for the AI model development is up to date, of high quality, complete and representative for the environment in which it is to be used (#2.3). To this end, ALTAI calls for the establishment of continuous measures (#4.1).

(III) Data preparation: During this phase, privacy (#3.1) and data governance (#3.2) aspects have to be considered. This includes, for example, access to data, which should be limited to authorized and qualified personnel (#3.1), and the integration and establishment of privacy-by-design measures. Only sources that comply with data protection laws or for

which the data owner has given consent should be selected for further processing (#3.1, #3.2). To comply with data governance, continuous monitoring measures should be defined and applied when constructing new data. Care should also be taken to ensure that the data selected is not biased and therefore has the potential to discriminate (#5.1). To achieve privacy-by-design and default, ALTAI recommends encryption, pseudonymization, aggregation and anonymization. Therefore, we propose the extension of the preprocessing phase to include a concrete processing activity for pseudonymization and anonymization. Care should also be taken to ensure that the data selected is not biased (#5.1). The same applies to subsequent activities where data are further processed. To avoid bias, continuous monitoring and measures such as simulation of new training data points for underrepresented classes should be included.

(IV) Modeling: Avoiding bias and the unfair discrimination of minorities (#5.1) is also a crucial requirement to be considered in the modeling phase. In concrete, for example, it should be checked whether a potentially suitable algorithm can deal well with unequally distributed data. When developing a test design, a testing activity that checks for discrimination should be explicitly integrated. For classification problems, for example, the use of the F1 difference - which relates the F1 score of the individual subclasses - is suitable [KFRE22]. Reproducibility and reliability also play an important role in modelling (#2.4). These factors must already be taken into account during model selection and integrated into a test concept. In addition to identify potentially harmful results generated by the AI model, it should also ensure that the training and execution of AI models is as resource-efficient and environmentally friendly as possible (#6.1). One measure for implementation could be, to consider the computational efficiency in the test design in addition to the pure quality of the results. Once the model training is complete, ensuring traceability (#4.1) and explainability (#4.2) of the model is important. Tracking should be provided of what data was used by the AI system to make a particular decision or in case of multiple models (stacked/ensemble architectures), which AI model or rules led to the results (#4.1). This property is closely related to the characteristic of model explainability. It should be ensured that both the technical processes of the AI system are explainable transparent as well as the reasoning behind the predictions of the AI algorithm.

(V) Evaluation: During evaluation the results of the model must be examined. Here, a variety of aspects and dimensions of ALTAI must be considered, as many of the aspects previously defined as requirements can be concretely evaluated based on the model results. For example, it should be specifically checked for discrimination using the previously defined performance metrics (#5.1). Another important aspect of the evaluation is to check the resilience of the AI model, since AI algorithms offer entirely new attack vectors. E.g. deep neural networks (DNNs) are vulnerable to adversarial examples added to the input by adding perturbations not noticeable by humans (#2.1). The evaluation should cover tests that check the behavior of the algorithm precisely for such cases.

(VI) Deployment: The application embedding the AI algorithm must be reliable, deliver reproducible results and include fallback scenarios (#2.4) and allow auditing by internal and external auditors (#7.1). Specifically, failsafe scenarios should be defined that are triggered, for example, in the case of results with low confidence or obvious errors. The system should be transparent to the end user that it uses an AI algorithm. This includes

information on the purpose, added values and risks and limitations of the AI service for the specific use case (#4.3). During development, potential external attacks on the AI system must also be considered and appropriate resilience mechanisms implemented (#2.1). Here, conformity to existing standards for cyber security and, if applicable, certification should be taken into account. This goes hand in hand with the consideration of general safety features. For example, the service should be designed to be resilient by using duplication (#2.2). Creating accessibility and adhering to universal design principles are central pillars of the deployment phase (#5.2). The general design of the AI system must correspond to the diversity of preferences and abilities in society. It is important to consider user groups with special needs and disabilities. Specifically, for example, it should be ensured that the interface is also accessible by users of assistive technologies. Users should be involved in a consultative manner throughout the implementation phase to ensure that the development team has an accurate understanding of their needs expectations and potential disadvantages caused by the system, e.g. explainability is often a crucial factor in increasing user confidence in AI systems (#4.2). When developing the user interface, an explanation function should be integrated that provides the user with as much information as possible about the decisions made by the system. If possible, this can be an explanation of why a model generated a particular output or decision and what combination of input factors contributed to it. The AI system should have minimal energy consumption and carbon emissions (#6.1) to ensure the well-being of the environment. A proper implementation should be based on the existing skills of the using employees, correspond to their knowledge and include self-explanatory functions (#6.2).

(VII) Operation: During operation, ongoing monitoring should be carried out to verify that the assumptions and database on which the modeling and training were founded are valid in practice and that the requirements for AI accuracy are met (#2.3). Such monitoring also serves to check the reliability (#2.4) and traceability (#4.1) of the AI system. A poor performance of the model can lead to unintended biases and thus to discrimination against individual groups (#5.1). To detect and prevent such misbehavior at an early stage, the previously selected fairness metrics should be continuously surveyed. In addition, the monitoring should also cover the resource consumption incurred for the model's operation (#6.1). If the AI system is to be used in a work context, the impact on employees and work processes must be recorded (#6.2). All monitoring aspects should be implemented in a way that promotes the auditability of the system (#7.1). Since ALTAI goes beyond mere monitoring and explicitly call for a feedback loop and adjustment of the AI system, we suggest extending the operation phase by the activity of a readjustment. In principle, all model behavior identified while monitoring and classified as undesirable should lead to an adjustment. Here the use of online learning, i.e., adjusting the model based on live data collected in the field, allows for better fitting the real data, but should take into account possible negative consequences in terms of the AI system learning unusual or undesirable patterns (#2.4). In addition to adjusting the AI algorithm, technical service updates may also be necessary to ensure the secure trustworthy execution, e.g., in terms of resistance to malicious attacks (#2.1). User interaction and communication are key to building trust in the AI system and a central component of ALTAI. Hence, we suggest the introduction of an activity that enables user interaction before, during and after the launch of the AI

system. This starts with simple communication about the duration of security coverage and updates (#2.1), but also includes informing users about the advantages potential risks associated with its use (#2.2, #4.3) and about the quality and accuracy of the predictions provided by it (#2.3). Enabling user interaction can help to increase their autonomy by making the origin of the systems output and functioning transparent thus promoting a more conscious decision-making (#1.1) This activity involves more than one-way communication, but also provides for the recording of and the response to user feedback. For example, the user must always have the option of withdrawing consent to the use of his or her data (#3.2). In this activity, measures should also be defined to ensure that the user understands. The success of these measures should then be evaluated (#4.2). In the case that the AI system itself communicates with the user (e.g. chatbots), it must be made transparent that this is an AI interacting with the user (#4.3). In addition, training material should be provided to enable users to adequately handle the system (#4.3). The material must also be usable by people with disabilities and by users of assistive technologies (#5.2). The user must also be given the opportunity to report observations related to bias, discrimination, or poor performance of the AI system (#5.1). In general, ALTAI advises broad inclusion of all potentially relevant stakeholder groups (#5.3), also during operation.

5 Discussion

While the EU guidelines itself are rather vague and make it difficult for companies to transfer them into practice, ALTAI offers concrete guiding questions and thus a stronger practical relevance. However, there is a lack of integration into practice-oriented procedure models. The model we have proposed is intended to remedy this situation and to help small and medium-sized enterprises in particular to implement ethical AI in practice. The mapping between activities and phases of the procedure model and ALTAI requirements ensures that the relevant guidelines are considered at all times in the development process. It can be noted that the ALTAI focuses on human users and their needs and rights. The perspective is also reflected in our process model and has an influence on how companies should develop AI models in the future to enable the realization of trustworthy solutions. Another relevant aspect is the continuity of measures required by the ALTAI in many places. Its implementation through mechanisms, such as MLOps, may lead to higher one-time investment costs for SMEs, but over time it can lead to the realization of efficiency gains and economies of scale that reduce the costs per AI project.

During the development of the procedure model, we had to decide on the level of detail of the model. The model had to be general enough to be independent of the industry and use case, and specific enough to ensure that the key activities were carried out. To ensure applicability, especially for SMEs with limited resources, the model should also be pragmatic and compact. Consequently we based it on the level of detail of CRISP-DM, since it is already widely used in practice and should also be familiar to SMEs. However, since our process model is an aggregate of various models, the implications can also be applied to them. This has the potential to simplify the integration of ALTAI, especially for

practitioners who already use procedure models and cannot easily switch from one to the other. Independent mapping by three experts and subsequent alignment is intended to strengthen reliability. To further increase the validity of the model, we want to encourage experts to apply and evaluate the model in practice. By developing the process model and integrating ALTAI, we answer the RQ with an instantiation of a procedure model that enables further research projects to integrate ALTAI requirements in the development process.

Every AI project is different and therefore, in practice, the focus on certain aspects of ALTAI must also be re-evaluated depending on the use case. Currently, our model focuses on stand-alone AI systems, and on systems in which AI is a core component. In the discipline of Information Systems, our work can be assigned to the development of prescriptive knowledge and can be used in Design Science research that intends to develop AI systems according to the EU guidelines. It also allows to draw conclusions about the completeness and suitability of existing AI systems for an integration of ethical requirements. We also found that existing and established models such as CRISP-DM do not sufficiently take ethical aspects into account by themselves and even existing extensions like Rebstadt et al. [RKGE22] are only addressing selected subparts like non-discrimination. Consequently, we suggest an expansion of the present models. In practice, our model created provides users with an introduction to compliance with the ethical guidelines, which in the future will become legally binding as a result of the EU's ambitions in the Artificial Intelligence Act and could result in penalties for companies that violate them. In this way, we want to reduce hurdles and uncertainties in the adoption of AI, especially for SMEs with limited resources. Beyond this introduction, the model can be further developed both scientifically and in practice. In addition to detailing the phases, it is also possible to expand the model to include additional best practices and tools in a general or domain specific manner. For example, we see great potential in embedding our process model in a software tool to enable SMEs in particular to use it even more concretely. Such a tool could be designed as an intelligent recommendation system that queries the user for information about the project over the individual phases and, on the basis of the input, makes concrete recommendations and points out necessary actions.

To date, AI is increasingly finding its way into very private areas and determines many decisions in people' s daily lives. At the same time taking ethical factors into account and creating trustworthy AI systems already goes far beyond mere compliance with laws such as GDPR and are likely to become even more important in the future. Although the ALTAI suggests performance tests - e.g. in terms of non-discrimination - they are rather vague regarding their concrete design. In our view, this results in a need for auditing and certification of AI systems based on standardized tests. Our work also has implications for the research field around data ecosystems. On the one hand, because AI systems are playing an increasingly important role here, and on the other hand, because data sharing, which is so central to data ecosystems, involves people. The data owners and producers are often private individuals whose willingness to share data with third parties is crucial to the success of ecosystems. By taking their interests into account more strongly during the development process and ensuring that their data is processed in trustworthy services, precisely this willingness could be strengthened.

6 Conclusion and Future Work

The increasing importance of developing and operating AI systems requires that fundamental rights are considered in the development process. To help developers of AI applications comply with the high ethical requirements, we have developed a procedure model that considers the EU guidelines for trustworthy AI applications. In a first step, we analyzed the existing procedure models in the area of ML around CRISP-DM and transferred them into an integrated model. Finally, by mapping ALTAI to the integrated model, we answer our research question and present our procedure model for a trustworthy development and operation of AI systems. Our model offers users an introduction to the holistic integration of measures to avoid discrimination, insufficient accuracy, lack of data protection and transparency in the development of AI systems and thus makes an important contribution to increasing trust in AI. Our proposed procedure model represents a basic building block, which will have to be extended in the future by more detailed investigations of the individual phases and methods. It offers SMEs in particular a starting point for the ethical use of AI. The intensive consideration of ethics in the development process and during the operation of AI systems can ensure that the great potential of AI applications can be used with reduced negative side effects in organization, private life and society. Future research in this area should continue the DSR cycle and focus on further evaluation of the model in practice and by experts (such as the High-Level Expert Group on AI). With this in mind, an evaluation by SMEs from various domains is planned for the future. In this context, we would also like to encourage other researchers and practitioners to apply the model to concrete AI projects and to contribute to its further development.

Bibliography

[ANIN17] ALTARTURI, HAMZA HUSSEIN ; NG, KENG-YAP ; IZUAN, MOHD ; NINGGAL, HAFEZ ; SHAHREL, AZREE ; NAZRI, AHMAD ; AZIM, ABDUL ; GHANI, ABD: A Requirement Engineering Model for Big Data Software. In: *Conference on Big Data and Analytics (ICBDA*, 2017 — ISBN 9781538607909, S. 111–117.

[BaDK20] BAUER, MARKUS ; VAN DINTHER, CLEMENS ; KIEFER, DANIEL: Machine learning in SME: an empirical study on enablers and success factors (2020).

[BaHN17] BAROCAS, SOLON ; HARDT, MORITZ ; NARAYANAN, ARVIND: Fairness in machine learning. In: *Nips tutorial* Bd. 1 (2017), S. 2017.

[CaCS20] CASTILLO, DANIELA ; CANHOTO, ANA ISABEL ; SAID, EMANUEL: The dark side of AI-powered service interactions: Exploring the process of co-destruction from the customer perspective. In: *The Service Industries Journal*, Taylor & Francis (2020), S. 1–26.

[CCKK00] CHAPMAN, PETE ; CLINTON, JULIAN ; KERBER, RANDY ; KHABAZA, THOMAS; REINARTZ, THOMAS ; SHEARER, COLIN ; WIRTH, RÜDIGER: CRISP-DM 1.0: Step-by-step data mining guide Bd. 16 (2000), S. 1–76.

[Euro21] EUROPEAN UNION, EU: Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 2021.

[FaPS96] FAYYAD, USAMA ; PIATETSKY-SHAPIRO, GREGORY ; SMYTH, PADHRAIC: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: *Communications of the ACM* Bd. 39 (1996), Nr. 11, S. 27–34.

[FeDe21] FERRÁNDIZ, ESTER MOCHOLÍ ; DEGLI-ESPOSTI, SARA: After the GDPR: Cybersecurity is the Elephant in the Artificial Intelligence Room. In: *European Business Law Review* Bd. 32 (2021), Nr. 1.

[High18] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, EUROPEAN COMMISSION: *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. Brüssel, 2018.

[HSMK19] HESENIUS, MARC ; SCHWENZFEIER, NILS ; MEYER, OLE ; KOOP, WILHELM ; GRUHN, VOLKER: Towards a software engineering process for developing data-driven applications. In: *Proceedings - 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering, RAISE 2019*, IEEE (2019), S. 35–41 — ISBN 9781728122724.

[JoIV19] JOBIN, ANNA ; IENCA, MARCELLO ; VAYENA, EFFY: Artificial Intelligence: the global landscape of ethics guidelines (2019).

[KFRE22] KORTUM, HENRIK ; FUKAS, PHILIPP ; REBSTADT, JONAS ; ELEKS, MARIAN ; NOBAKHT GALEHPARDSARI, MARJAN ; THOMAS, OLIVER: Proposing a Roadmap for Designing Non-Discriminatory ML Services: Preliminary Results from a Design Science Research Project (2022).

[KoGe20] KOMMISSION, EUROPÄISCHE ; GENERALDIREKTION KOMMUNIKATIONSNETZE, INHALTE UND TECHNOLOGIEN: *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment* : Publications Office, 2020.

[KoRT22] KORTUM, HENRIK ; REBSTADT, JONAS ; THOMAS, OLIVER: Dissection of AI Job Advertisements: A Text Mining-based Analysis of Employee Skills in the Disciplines Computer Vision and Natural Language Processing. In: *Proceedings of the 55th Hawaii International Conference on System*

Sciences, 2022.

[LaRa11] LAKSHMI, B. N. ; RAGHUNANDHAN, G. H.: A conceptual overview of data mining. In: *Proceedings of National Conference on Innovations in Emerging Technology, NCOIET'11*, IEEE (2011), S. 27–32 — ISBN 9781612848082.

[LeLL21] LEBOVITZ, SARAH ; LEVINA, NATALIA ; LIFSHITZ-ASSAF, HILA: Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. In: *Management Information Systems Quarterly* (2021).

[MaMF10] MARISCAL, GONZALO ; MARBAN, OSCAR ; FERNANDEZ, COVADONGA: A survey of data mining and knowledge discovery process models and methodologies. In: *The Knowledge Engineering Review* Bd. 25, Cambridge University Press (2010), Nr. 2, S. 137–166.

[MCFH19] MARTINEZ-PLUMED, FERNANDO ; CONTRERAS-OCHANDO, LIDIA ; FERRI, CESAR ; HERNANDEZ ORALLO, JOSE ; KULL, MEELIS ; LACHICHE, NICOLAS ; RAMIREZ QUINTANA, MARIA JOSE ; FLACH, PETER A.: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. In: *IEEE Transactions on Knowledge and Data Engineering* Bd. 4347 (2019), Nr. c, S. 1–1.

[MiCo19] MILLER, C ; COLDICOTT, R: People, power and technology: The tech workers' view. In: *Retrieved from Doteveryone website: https://doteveryone. org. uk/report/workersview* (2019).

[RKGE22] REBSTADT, JONAS ; KORTUM, HENRIK ; GRAVEMEIER, LAURA SOPHIE ; EBERHARDT, BIRGID ; THOMAS, OLIVER: Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services. In: *HMD Praxis der Wirtschaftsinformatik* Bd. 59 (2022), Nr. 2, S. 495–511.

[RoBL20] ROBERT JR, LIONEL P ; BANSAL, GAURAV ; LUTGE, CHRISTOPH: ICIS 2019 SIGHCI workshop panel report: human--computer interaction challenges and opportunities for fair, trustworthy and ethical artificial intelligence. In: *AIS Transactions on Human-Computer Interaction* Bd. 12 (2020), Nr. 2, S. 96–108.

[Royc87] ROYCE, WINSTON W: Managing the development of large software systems: concepts and techniques. In: *Proceedings of the 9th international conference on Software Engineering. IEEE Computer Society Press* (1987).

[SaGr17] SALTZ, JEFFREY S. ; GRADY, NANCY W.: The ambiguity of data science team roles and the need for a data science workforce framework. In: *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017* Bd. 2018-Janua (2017), S. 2355–2361 — ISBN 9781538627143.

[SeBM21] SEPPÄLÄ, AKSELI ; BIRKSTEDT, TEEMU ; MÄNTYMÄKI, MATTI: From Ethical AI Principles to Governed AI. In: *ICIS 2021 Proceedings* (2021).

[SiSS19] SILVA, CHARITH ; SARAEE, MAHSA ; SARAEE, MO: Data Science in Public Mental Health: A New Analytic Framework. In: *Proceedings - IEEE Symposium on Computers and Communications* Bd. 2019-June (2019), S. 1123–1128 — ISBN 9781728129990.

[TiOH21] TIMAN, TJERK ; VAN OIRSOUW, CHARLOTTE ; HOEKSTRA, MARISSA: The Role of Data Regulation in Shaping AI: An Overview of Challenges and Recommendations for SMEs. In: *The Elements of Big Data Value*, Springer, Cham (2021), S. 355.

[VaWi18] VANDERELST, DIETER ; WINFIELD, ALAN: The dark side of ethical robots. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, S. 317–322.

[Wang08] WANG, XIAOH: The realization of knowledge discovery in total quality management system. In: *Proceedings - 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008* Bd. 4 (2008), S. 643–646 — ISBN 9780769533056.

[WaSi18] WANG, WEIYU ; SIAU, KENG: Ethical and moral issues with AI: a case study on healthcare robots. In: *Twenty-fourth Americas conference on information systems. Retrieved from July*. Bd. 16, 2018, S. 2019.

[WeWa02] WEBSTER, JANE ; WATSON, RICHARD T: Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: *Management Information Systems Quarterly* Bd. 26 (2002), Nr. 2, S. xiii–xxiii — ISBN 02767783.

[WoMG17] WOLF, M J ; MILLER, K W ; GRODZINSKY, F S: Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. In: *The ORBIT Journal* Bd. 1 (2017), Nr. 2, S. 1–12.

[YHZL19] YUAN, XIAOYONG ; HE, PAN ; ZHU, QILE ; LI, XIAOLIN: Adversarial examples: Attacks and defenses for deep learning. In: *IEEE transactions on neural networks and learning systems* Bd. 30, IEEE (2019), Nr. 9, S. 2805–2824.