Validation of AI-based Information Systems for Sensitive Use Cases: Using an XAI Approach in Pharmaceutical Engineering

Polzer, Anna Katharina
University of Graz
anna.polzer@uni-graz.at

Fleiß, Jürgen
University of Graz
juergen.fleiss@uni-graz.at

Ebner, Thomas
KML Vision GmbH
thomas.ebner@kmlvision.com

Kainz, Philipp
KML Vision GmbH
philipp.kainz@kmlvision.com

Koeth, Christoph
Fresenius Kabi Austria GmbH,
Christoph.Koeth@fresenius-kabi.com

Thalmann, Stefan
University of Graz
stefan.thalmann@uni-graz.at

Abstract

Artificial Intelligence (AI) is adopted in many businesses. However, adoption lacks behind for use cases with regulatory or compliance requirements, as validation and auditing of AI is still unresolved. AI's opaqueness (i.e., "black box") makes the validation challenging for auditors. Explainable AI (XAI) is the proposed technical countermeasure that can support validation and auditing of AI. We developed an XAI based validation approach for AI in sensitive use cases that facilitates the understanding of the system's behaviour. We conducted a case study in pharmaceutical manufacturing where strict regulatory requirements are present. The validation approach and an XAI prototype were developed through multiple workshops and then tested and evaluated with interviews. Our approach proved suitable to collect the required evidence for a software validation, but requires additional efforts compared to a traditional software validation. AI validation is an iterative process and clear regulations and guidelines are needed.

1. Introduction

Advances of Machine learning (ML) / Artificial Intelligence (AI) have led to their increasing use. However, while a particularly high adoption of such technologies can be observed in some areas like online marketing, high-risk or other sensitive industries that are highly regulated like banking or finance show a low adoption of AI [1]. The latter often have higher standards for proper quality assurance/validation when adopting new technologies [2]. New technology needs to be audited or validated before it can be used. However, auditing and validating of AI is challenging [3, 4]. This is a major barrier for adoption of AI-based applications in sensitive use cases [5].

AI systems need to be evaluated for possible negative consequences of inadequate functioning to minimize risks. While interpretable ML models are (to some degree) directly comprehensible [6], many AI systems rely on non-transparent ML algorithms, i.e., black box algorithms [1]. For the latter, the limited understanding of the decision process is considered a major drawback when implementing them [7]. Examples of biased AI systems [8] furthermore increase caution when considering their use in sensitive areas as decisions based on biased data or wrong decision criteria can have major consequences. Thus, such systems need thorough validation that properly addresses AI specific issues like the black box problem. However, practical guidelines on how to validate AI systems are still missing. The scientific literature suggests that advances in the field of explainable artificial intelligence (XAI) will be instrumental in providing insights about black box AI systems [9, 10]. Understanding the underlying decision rules might allow some confidence in the forecast of an AI's behaviour in new and previously unknown situations, which is essential when the consequences of errors are potentially severe. However, most of the existing work lacks empirical evidence gathered through studies of existing XAI frameworks and their ability to satisfy the needs of stakeholders [2]. Many authors suggest that XAI might be helpful in the validation [2, 6, 9, 10], but specific approaches have yet to be developed and evaluated. At the same time, issues concerning the auditing and validation of AI systems are becoming a more prominent topic [4, 11].

We aim to combine these two actively discussed aspects in AI system research and focus on the following research question: How can XAI be utilized to validate AI systems in sensitive use cases? To answer this question, an XAI based validation approach for AI in sensitive use cases was developed by a team of academics, industry experts and an auditor. In a

URI: https://hdl.handle.net/10125/79518
978-0-9981331-5-7

case study at an Austrian pharmaceutical manufacturing company we evaluated whether the proposed approach gathers sufficient information for a successful validation of an AI system in this sensitive use case. The results show that XAI is essential for a successful validation of AI systems in the highly regulated pharmaceutical industry. In the following sections, we will first discuss the related work before introducing the case study. Next, the proposed validation approach is presented as well as the underlying AI system and applied XAI approach. Finally, we reflect on the evaluation results of this case study. In conclusion, we discuss the results and theoretical and practical implications.

2. Background

Recent developments in AI technologies increased the performance of AI based systems [5]. However, at the same time they often become increasingly more complex. With rising complexity, it is getting more difficult to extract and understand the underlying reasoning. This circumstance is often called the "black box problem" [5]. The black box character of AI is often considered a primary concern for trust in and acceptance of AI systems [9, 12].

Validation in highly regulated areas- Understanding the inner workings of AI systems is not always needed in cases where the consequences of failed decisions are less severe [12]. However, the black box character can be a limiting if not disqualifying factor in critical applications [5]. In safety critical and thus usually highly regulated industries like the pharmaceutical sector, it is mandatory to assure the reliability of introduced technologies prior to integration into operating processes [13, 14]. Different procedures of assuring suitability and reliability of information systems (IS) in business processes can be found, e.g., quality assurance, software verification, software validation or IT auditing [15]. The common goal is to systematically compile evidence to ensure consistency/reliability and compliance with the specifications of the IS. Aligned with the terminology of regulations relevant in the pharmaceutical industry [13, 14], we use the term validation to subsume the process of providing *"[...] objective evidence, that the requirements for a specific intended use or application have been fulfilled"*[15].

Following the prescribed validation of software, several official guidelines like the FDA'S *"General principles of Software Validation"*[16] or the European Commission's *"Good Manufacturing Practice (GMP) guidelines Annex 11: Computerized Systems"*[17] as well as industry specific approaches were developed in the recent decades [18]. Following the regulatory guidelines, the approaches are predominantly linear and sequential. However, existing software validation approaches cannot be carried over to present AI systems as they face several open issues.

Open Issues- A fundamental difference of AI systems compared to traditional software is the importance of data in the development process. In traditional software all application logic and instructions are expressed by the developers in source code [11]. AI systems based on ML learn their application logic independently of developers from provided training data. The importance of data quality is thus a pivotal factor, but clear guidelines are still missing [19, 1]. Additionally, as the ML research (which dominates contemporary AI systems [11]) is still in its early stages, many fundamental issues like terminology and standardized guidelines for ML quality are missing [11, 20]. A recent upturn in ML testing research has been noticed [21], however issues like lack of specification and defined requirements in learning-based approaches or the interpretability of black box models/systems are still to be determined [20].

As there are many open issues and no way to verify correctness of AI systems by mathematical proofs, Winter et al. suggest to *"[...] validate whether a ML approach is reasonable, correct, meaningful and clear"* [4]. Depending on the effects of the application's decision on people, environment, and organizations they assign criticality levels. The higher the criticality level, the more extensive the testing for the ML application has to be. In more critical applications, the audit catalogue demands explainability and interpretability [4].

XAI as facilitator for the validation of AI systems- Explanations for an AI system's behaviour can be provided trough XAI approaches. XAI makes AI systems understandable to humans [9]. Many XAI techniques have been developed to enable post-hoc explanations for opaque ML models. They enhance interpretability using textual or visual explanations, explanations by example or feature relevance explanations [6]. XAI approaches for computer vision are developed for specific ML models. Because artificial neural networks (NN), especially convolutional NN, are often used in computer vision, many model-specific XAI approaches like heatmaps [22] and class activation methods (GradCAM) [23] were developed, too. There are also prominent model-agnostic approaches like LIME [24] and SHAP [25] that can be used in AI-based computer vision systems. A comprehensive review of XAI approaches was recently conducted by Arrieta et al. [6]. XAI approaches can reveal hidden features influencing the decisions made by AI. Including XAI in the validation of AI systems can help detect underlying problems which could not have been detected by only focusing on conventional evaluation metrics [26]. Furthermore, the analysis of different ML models developed for the same purpose examined with a visual XAI approach indicated that there might be significant differences in the explanation provided by the ML models [7]. While it has been shown that XAI can provide insights into black box AI systems in various contexts [24, 26], to date there is no systematic approach on how to utilize XAI to assure that the developed AI systems act reliably. To address this research gap, this paper proposes a validation approach that addresses all necessary information needs for a validation in a highly sensitive area like the pharmaceutical industry.

3. Case Study

The pharmaceutical industry is known for its need of high regulatory compliance. In this context, every newly introduced IS or process has to go through rigorous validation procedures to assure its adequate functioning. Existing and established regulations are tailored to validation of classic soft- and hardware [13, 17, 27]. AI based systems are simply not covered. This is known and attempts to resolve that shortcoming were started [28]. Nevertheless, a significant hesitation in implementing new AI systems arises from the uncertainty concerning the AI systems' decision behaviour.

Our case study is about an Austrian manufacturing plant for parenterals (sterile pharmaceuticals). Sterile drug manufacturing requires strict adherence to operating procedures and timely documentation of all activities, especially in the sterile core where product is filled into vials. Although the open product is protected and separated from human operators at all times, interventions above the open product may impose the risk of introducing particles and results in rejection of possible affected products. Even for experienced operators, it is sometimes difficult to make a precise distinction between critical and non-critical interventions (and thus about the necessary corrective actions). It is only understandable that human interventions have to be timely documented, classified and analysed in order to produce the highest possible quality. Precisely this contradiction - sterile operation without humans is not feasible currently, but humans impose the highest risk - can be mitigated through the use of an AI system.

The proposed AI based system is set to detect the time and place of human intervention in an otherwise (machine) automated process, and to additionally classify whether the intervention should result in the removal of possibly affected vials present during the intervention. This case study focuses on the issue of understandability and interpretability of AI models and how the introduction of XAI component(s) in the validation procedure might help in facilitating trust and provide evidence of the AI system's adequate behaviour to be implemented in a highly sensitive area.

3.1. Procedure

To assure the proper functioning and results of the AI system a validation approach is needed that better addresses the specific issues emerging from AI technologies. As validation in critical use cases needs to consider all possible situations, testing the performance with a set of test cases is not sufficient. Understanding the underlying decision behaviour of the AI system is necessary to uncover wrong decision criteria that could lead to negative consequences in certain settings. For example, consider that the presence of a stationary object in the production line is learned as a main decision criterion. However, in reality this component is irrelevant for the differentiation between critical and non-critical interventions, but simply results from biased training data where the component was present in all critical interventions. Black box AI systems would not allow to understand that the presence of this object influences the decision. Thus, the overarching goal of the presented collaborative research and development project was to develop a validation concept for AI that would address and satisfy the information and explanation needs during a software validation process in the pharmaceutical industry. A team of academic researchers, industry and technology experts as well as a GMP auditor conducted ten requirement workshops over the time span of seven months. In each workshop the experts of their field addressed specific issues regarding the situation at the pharmaceutical company, regulatory restrictions and technical possibilities which were then discussed and analysed. Concurrently, a literature review was conducted to map the challenges and approaches to validate AI systems identified in scientific literature. The result was a collection of issues and requirements for the AI system and its validation concept that needed to be addressed adequately. Expanding on the requirements from traditional software validation the main AI specific issues that needed to be adequately addressed were data and transparency of decision mechanisms.

Based on this compilation of requirements, a preliminary validation concept was proposed. The main goal was to increase transparency and provide evidence suitable for a software validation in the pharmaceutical industry. The drafted validation concept was then iteratively adapted until it was conceived satisfactory and implementable by all involved parties. Concurrently, the AI system with its XAI component was developed to pose as a prototype to be used for the developed validation concept.

3.2. Learning and explaining criticality of interventions from video streams

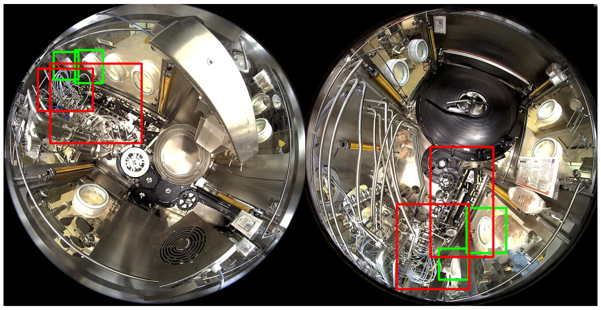

Figure 1. Example input frame showing two corresponding glove ports from two fixed camera positions. Green regions indicate HOG sampling areas for detecting an intervention for a specific port, red regions for classification of the criticality

To detect interventions, i.e. glove insertions, two fixed wide-angle cameras located at the ceiling of the aseptic core (”isolator”) are used to capture a continuous video stream at 10 frames per second. We formulated the task as ML problem and implemented a two-stage computer vision algorithm to analyse the video stream in real-time.

For each glove port p, a Random Forest (RF) [29] algorithm first learns a binary classifier C^p_{int} of whether the glove is inside (positive) or outside the isolator (negative). Each frame is classified independently based on Histogram of Oriented Gradient (HOG) features [30], computed within a manually defined region for each glove port, see green regions in Figure 1. For this ML problem, this particular choice of classifier and features achieved more robust results than state-of-the-art NNs, which tended to over-fit to our limited number of training samples. The training set consists of positive samples from actual glove insertion sequences and negative samples from other sequences, e.g. where operators moved in the vicinity of the glove port, or performed insertions in neighbouring glove ports.

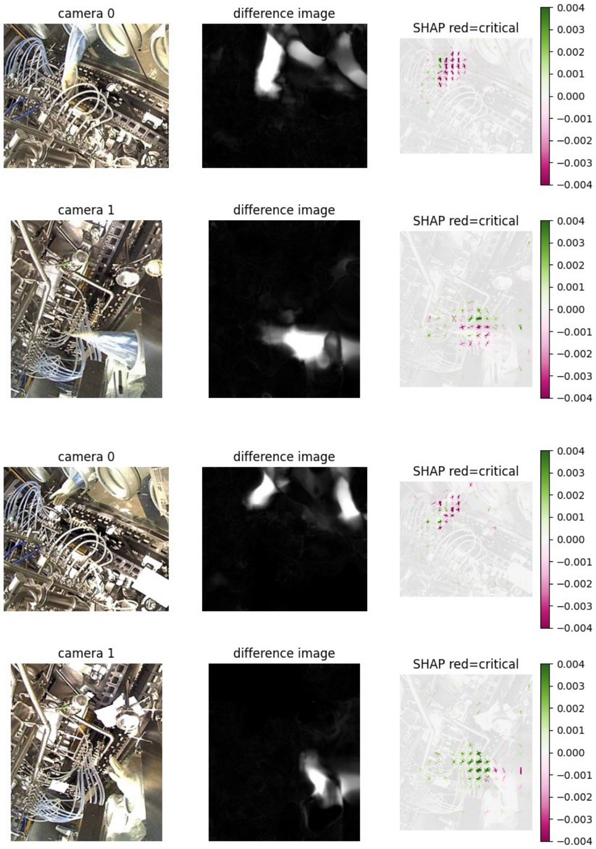

Figure 2. Visualization of SHAP values for HOG features significant to the criticality decision. Critical (top) and non-critical (bottom) frame for a glove position from both cameras, masked by the difference image indicating movements

To distinguish between critical and non-critical interventions, a second binary RF classifier C^p_{crit} is trained, again for each glove port. It uses HOG features from a larger region around each glove port (red regions in Figure 1), to assess the criticality of each frame, where C^p_{int} detects the glove inside. HOG features are computed on the difference image between the current and the first frame before the intervention. Once a single frame is detected as critical, the whole intervention is considered critical. This training set consists of critical and non-critical glove insertion sequences. Critical sequences also contain non-critical frames, typically in the beginning and at the end, and at least one critical frame. Since the frames do not contain ground truth criticality labels, we use a pre-training stage that automatically determines, which frames are critical and which frames are not critical. These generated per-frame labels are then used to build C^p_{crit}. Data for supervised pre-training was generated by a weak-label heuristic: all frames of critical sequences, except the first and last n frames, were labelled as critical, the remaining frames as non-critical. Depth-reduced RFs C_{sel} were trained on HOG features to prevent over-fitting to this partially wrong labels. In a subsequent step, we predict the criticality probability using the trained C_{sel} on all training samples. Critical frames were selected as those with a probability above 70% of the maximum probability within each critical sequence. All remaining frames were considered as non-critical and finally, C^p_{crit} was built on this data set.

To allow humans to gain insights into why the RF classified sequences as critical or non-critical, we applied SHapely Additive exPlanations (SHAP) [25] to visualize HOG features used by C^p_{crit} in an image. Figure 2 shows positive SHAP values (contribute to decision ”non-critical”) in green and negative SHAP values (contribute to decision ”critical”) in red. Evaluating the AI systems through the lens of XAI enables the assurance that the decision spaces match with actual critical interventions in the manufacturing process. In this case study, SHAP was chosen as the appropriate XAI tool, because of its theoretically sound foundation and also its model-agnostic nature. Many of the current XAI tools for computer vision task are specifically developed for NN. Moreover, as the underlying ML model is a random forest, SHAP was chosen to be the most suitable XAI approach.

3.3. Evaluation

To gather insights on the applicability and possible obstacles in this validation approach, a group of employees of the pharmaceutical company and a GMP auditor performed the proposed validation approach for the developed AI system. Afterwards they were interviewed and asked about their experience.

Data collection and participants - A group of four employees of the pharmaceutical company and an external participated in the evaluation: (1) the director of innovation of the pharmaceutical company, (2) the teamlead of operators of the sterile filling line the AI system was implemented for, (3) one from the manufacturing quality assurance (MQA) department, (4) and one responsible for the validation of aseptic production processes, and (5) an external GMP auditor.

Semi-structured interviews to gather their experience in and opinions on the validation approach were conducted after its completion. Employees directly responsible for implementing the validation process (i.e., interviewees 2-4) were interviewed twice in group interviews: Once after testing the AI system without the XAI component and a second time after completing the whole validation approach including the XAI component. Special focus in those interviews was put on the perceived differences between the first half of the procedure, in which they solely examined the performance of the system as a black box, and the second half where the XAI component is included. Interviewee 1 and 5 were interviewed separately and only once on the whole validation approach.

Data analysis - The interviews, conducted via web-conference, were audio recorded and transcribed. The transcripts were analysed using semantic thematic analysis [31] to identify and organize barriers as well as positive aspects of the evaluation process in general, and with regard to the XAI component in particular.

4. AI Software Validation approach

The validation concept consists of four phases which may be completed step by step or repeated several times. The first phase – *Set-up phase* – deals with typical design tasks of developing and implementing a new IS. The following *Technical set-up and performance* phase addresses AI specific development steps for developing a reliable AI system. Up to this point the procedure is similar to validation approaches or performance evaluations of black box AI systems. While predefined performance measures may be sufficient for AI systems in non-sensitive areas, the proposed validation concept takes additional efforts to validate the developed AI system and to ensure that it acts as intended. The third *XAI Assessment* phase introduces the proposed XAI approach showing the appropriate use of decision rules and/or spaces. Building on this assessment, the last phase, the *Stress test*, attempts to test the limits of the AI system's applicability.

4.1. Set-up

The *Set-up phase* focuses on initial design and planning activities. Following an idea for a new application of an AI system the *Set-up phase* starts with traditional business considerations in mind. In the first step - *Use case definition* - all initial planning activities take place. Here, the project is set up. The initial requirements for the AI system, necessary further planning and implementation steps as well as responsibilities are defined. When the IS's concept is properly analysed and considered worthwhile for further development, suitable business units, employees, experts and/or third-party suppliers for further project steps need to be determined.

Form expert team - Based on the use case definition, the expert team should be formed. It will be responsible for all validation steps and thus needs to have the proper knowledge and authorization. The team should include domain experts, end-users, quality assurance, project management and technology

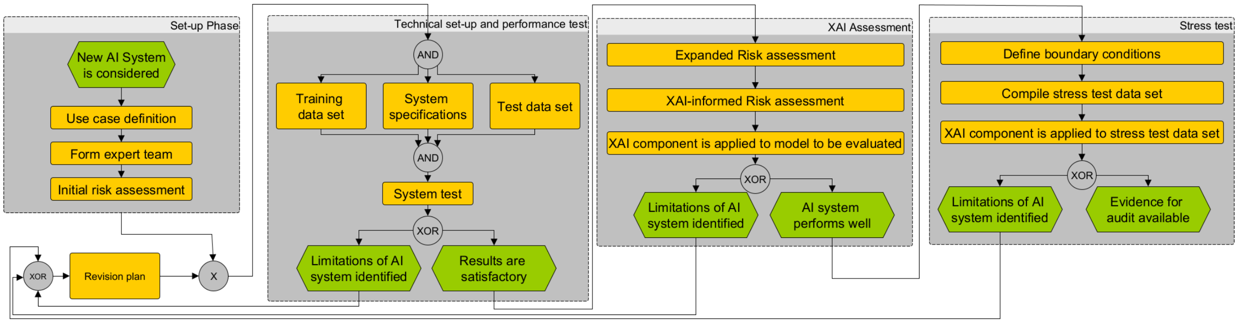
Set-up Phase
New AI System is considered
Use case definition
Form expert team
Initial risk assessment
XOR
Revision plan
X
Technical set-up and performance test
AND
Training data set
System specifications
Test data set
AND
System test
XOR
Limitations of AI system identified
Results are satisfactory
XAI Assessment
Expanded Risk assessment
XAI-informed Risk assessment
XAI component is applied to model to be evaluated
XOR
Limitations of AI system identified
AI system performs well
Stress test
Define boundary conditions
Compile stress test data set
XAI component is applied to stress test data set
XOR
Limitations of AI system identified
Evidence for audit available

Figure 3. Developed validation approach

suppliers. As technology suppliers bring expertise into the development from a technical perspective, domain experts provide domain knowledge that is necessary to determine the requirements and functionalities of the specific AI system. For the assurance of the proper validation and subsequent achievement of quality standards of the AI systems, quality assurance (QA)/auditing experts are needed to provide guidance in pinpointing critical aspects of the validation process.

Initial risk assessment - The expert team conducts an initial risk assessment for the implementation of the AI system. This can follow different methods according to the company's policies or the specific situation. Depending on the criticality of the use case, the risk assessment should cover all aspects that could be affected by the implementation of the new AI system including business and quality issues, technical and AI specific risks, risks to the end-users and risk to society.

4.2. Technical Set-up and Performance

System specification – This step occurs concurrently with the definition of the training and test data set for the planned AI system grouped into the *Technical Set-up and Performance Test* phase. In this step, the requirements are investigated and appropriate system specifications are agreed upon. Among these are decisions about the features of the AI, the ML approach used to satisfy these requirements, and the performance metrics. This is a step where especially the expertise of domain-experts and quality assurance comes into play. Depending on the context of the use case possible industry standards on acceptance criteria etc. might exist that need to be followed. In this step the XAI component needs to be specified. Depending on the use case, a suitable decision has to be made about which XAI component will be applied and how it will deliver the explanations needed for the validation.

Training data set - The step of the compilation of the training data set concurs with the elaboration of the system specifications. To ensure that the training data set represents everyday processes (which the AI system must be able to handle appropriately) a team of domain experts and technology providers need to elaborate a scheme on how to compile a satisfactory training data set. The scheme must consider traditional data issues such as correct distribution, labelling, potential biases, completeness and sources as well as domain specific issues that may arise in the context in which the AI system is going to be used. According to this scheme the training data set needs to be created (and documented).

Test data set - Defining the test data set is a crucial task. It needs to adequately represent the real-life challenges the AI system is intended for. It should include as many scenarios as possible which are likely to happen under regular circumstances. Furthermore, it has to be assured that the distinction between the training data set and the test data set has been handled adequately to avoid overestimation of the systems performance.

System test – The system test serves as a basic performance test of the developed system to assure it meets the required targets. It's tested against the system specifications and the results are compared with the predefined acceptance criteria. Depending on the results, the system is either send back into revision or ascends into the next phase of the validation concept.

4.3. XAI Assessment

Expanded Risk Assessment – If the results of the performance test are satisfying, the initial risk assessment gets re-evaluated and expanded according to the new information gathered in the *Technical set-up and performance test phase*. Specific decisions made in this phase may lead to additional risks that were or could not have been considered beforehand. Again, in this step possible mitigation approaches for the identified risks may be defined and decided upon.

XAI-informed Risk Assessment – In a subsequent step a risk assessment is conducted that specifically analyses the risks associated with the black box characteristics of the developed AI system. It should investigate risks of the AI system using non-relevant/trivial information/data as decisive factors for its decisions and weigh the possible consequences of the system showing such behaviour. For instance, the AI system might classify an intervention as critical solely because of the presence of a non-stationary component in the assembly line and not consider the actual intervention space as the crucial area for its classification. The results of this step should provide a round-up of critical areas/decision spaces the AI systems should comply with to be considered appropriate for its purpose.

XAI component is applied to the model to be evaluated - After identifying the critical decision spaces of the AI, the XAI component is applied to the results of the system test where it pinpoints the features/areas considered for the decision taken by the AI. The expert team needs to assess if the areas shown comply with the defined decision criteria and are both meaningful and suitable for each particular decision. If the decision-making of the AI is satisfactory, based on the XAI evidence, the validation procedure continues, otherwise the AI system is sent back to be revised.

4.4. Stress Test

Define boundary conditions - After the XAI assessment phase has concluded, the final validation stage of the AI begins. Based on the evidence from the previous phases the expert team identifies critical areas and defines possible boundary conditions. These boundary cases consist of unlikely but possible scenarios for the AI system and are established to evaluate whether the system behaviour matches the requirements for specific boundary conditions. Considerations about the inclusion of adversarial attacks, random interventions etc. can be included.

Compile stress test data set - After defining the boundary conditions, necessary data has to be compiled. Real processes data are preferable, but synthetic data are acceptable if that is not possible or unsafe.

XAI component is applied to stress test data set – As a last step the XAI component is applied to the stress test data set. This step serves as a final examination and assurance of the limits of the AI system and whether they correspond to the requirements of the specific use case. Similar to the steps in the *Risk Assessment phase*, the stress test data acts as the input to the AI model and the results are then investigated with the help of the XAI component. If the AI system behaves as expected, the results can be documented and evidence for an audit is available. In the case that limitations of the AI system are detected, the AI system needs to be revised.

Revision plan - The revision plan depends on the identified limitations. It might include minor adjustments like adding a particular instance to the training data or big adaptations like changing specifications or even the underlying ML algorithm. Furthermore, this step can be the starting point of iterations in the development process of the AI system or might not come into effect at all.

Although some steps of the developed validation approach by themselves might not be considered novel, a systematic approach on how to integrate XAI in AI validation endeavours has not yet been examined for sensitive areas. The presented approach provides a general guideline on how complex and opaque AI systems might be validated to achieve the required level of transparency in critical use cases. As the validation approach does not specify underlying ML models of an AI system or specific XAI tools, the approach can easily be adapted to be used in different environments, even outside critical use cases.

5. Discussion of evaluation results

The validation approach that is evaluated in this section was carried out in practice for the AI system presented above. A total of 213 interventions into the aseptic core were recorded and evaluated. These interventions were in the first step evaluated according to the system test in phase two. In the next step, the same set of interventions was examined in the XAI assessment phase. The employees of the pharmaceutical company examined the whole data set including all correctly as well as all incorrectly classified interventions.

Phase One- Set-up Phase Regarding traditional and AI validation approaches the director of innovation stated that *"[i]n this phase it is possible that people say 'Well there are not really any differences between traditional validation and this approach'. [... T]here are not many options to do this differently"*. Others also did not consider this phase particularly different compared to traditional software validation. Some particularities in the use of AI were mentioned as needing to be considered during the *Use Case Definition* and the *Initial Risk Assessment*. Those were the precise and suitable formulation of requirements or acceptance criteria, but also consideration of AI-specific risks. The GMP auditor noted that the steps might not be simply linear, but may resemble *"[...] more of a cycle (iterative process), where the steps blend into each other"*. Specific risks might emerge with a particular AI component and that again can influence the requirements of the AI system. Extensive communication on the requirements and limitations of the AI systems is also seen as necessary [4]. The GMP auditor also argues that standards or guidelines from regulatory bodies would benefit the set-up phase.

Phase Two- Technical Set-up and Performance In phase two, differences between traditional validation approaches and our proposed approach were identified. Citing the need for high quality training data, interviewees agreed that traditional validation approaches need to be adapted to include this AI specific issue: Issues of data and data handling have to be incorporated much more prominently. In this regard the documentation of AI differs compared to traditional software, especially for the documentation of training and test data sets and the performed data preparation. However, documentation standards for AI are missing and more research is needed [1].

Interviewees highlight that testing of AI based systems and traditional software differs. The director of innovation points out that "*[...] traditional qualification or software validation only asks the question 'does it work or not'. Here you do have to think about it in more detail"*. With traditional software, decision rules are explicitly implemented and test cases designed accordingly. For AI to "think in more detail" entails uncovering the decision rules and to design boundary cases for testing. Similar concerns regarding AI testing are also raised in the literature [20, 21]. The integration of AI systems in business processes was identified as another challenge. Several interviewees considered it a significant issue to make tacit knowledge about existing processes explicit so it can be included in the requirements specification and the compilation of training and test data. However, working through these issues resulted in a significantly deeper understanding of existing processes as well as better knowledge on how to transfer the insights gathered to the AI system. The need for standards or guidelines for the validation of AI was highlighted in this phase.

Phase Three- XAI Assessment The application of the XAI component was seen as important to create trust in the AI system. The XAI component, e.g., helped to detect bias in the training data set. Early versions of the AI system decided on the criticality of an intervention based on the presence of a second hand inserted into the isolator (Figure 4). In the training data set one-handed interventions were used to depict critical interventions, while two-handed interventions mostly depicted non-critical interventions. Traditional quantitative performance metrics were still relatively good.

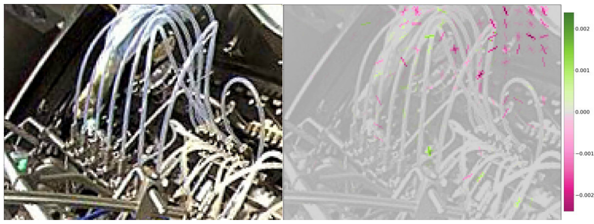

Figure 4. Example of detected data bias using the XAI component. The AI was focusing on the second hand (upper right corner) to determine criticality

Based on such experiences our interviewees considered this approach a good way to make the decision process of the AI comprehensible. As the director of innovation noted with regard to the example above: *"We would have never been able to recognize it using a black box approach. It would probably be impossible."* This is a specific example of the generally accepted view that without XAI it is not possible to comprehend the deeper underlying reasoning of many AI systems [26]. Another useful result of using XAI was that in some cases it also allowed to counter-check if the decision spaces and criteria defined in the systems specification steps in phase two were set precisely enough.

Interviewees agreed that XAI helped to understand the AI system by making the inner workings of the AI transparent. As the MQA representative put it: *"In normal validations [non AI] you would see that the software decides incorrectly, but one does not really know why. However, by having it processed on this level of detail for this AI system, everything becomes more comprehensible"*. XAI was also considered as beneficial for recognizing errors or oversights that might have happened in the previous steps.

Phase Four- Stress Test Interviewees considered the definition of boundary areas and the examinations on whether the AI system would work in the limits of these boundaries to be sound. It was remarked that it seemed a good approach to identify boundaries not just based on the initial requirements, but also to derive such boundary areas or condition from the specific implementation of the AI system itself. This possible adjustment of requirements based on the findings from the XAI evaluation might be a possible countermeasure for the often cited lack of specifications and defined requirements in the beginning of a development process [20]. It also takes into consideration that the explanations on the decision areas from different AI systems may differ [7]. Thus, the setting of the boundaries is done after a first examination with the XAI component was conducted.

The evidence showing that the AI systems works reliably in the defined boundaries builds trust in the application. As the teamlead of operators said: *"Trust in a system is always necessary, but especially in our sector it is necessary to have provable evidence. In my opinion this approach that we just executed with the XAI component, does this well."* The MQA employee remarked that the XAI component delivers insights in cases where humans could have difficulty classifying an intervention where the technical camera set-up does not always provide the most suitable perspective for human review. Additionally, it was again remarked that such tests on whether it is even possible to achieve not just performance but also behaviour goals through the use of an XAI component were happening relatively late in the validation procedure. This further strengthens the argument that a more agile development and validation approach is preferable for AI.

Concluding remarks The final question on whether one would trust to deploy an AI system that was validated according to the proposed validation approach was unanimously answered positively. The additional effort required by the proposed approach not only led to new insights into the behaviour of the AI, but also to new insights about the implemented case (especially about the boundary cases) and the production process in general. The implementation of the sufficiently validated AI system also enables a continuous quality assurance of the production process that otherwise would not be feasible.

Implication for research and practice The evaluation of the validation approach showed that XAI can be successfully utilized to support validation efforts in sensitive areas. However, the repeated mentioning of missing standards and regulations indicates that there is still a lot of uncertainty for all parties involved. As there are no clear criteria for the amount of data and their quality, criteria have to be defined by those examining the AI system (i.e., internal validation, external auditor, certification authority). Further, the evaluation results showcase that a more agile validation approach for AI is preferable and that validation considerations should be included early in the process. The developed validation approach encourages this iterative approach by incorporating possible loops after every phase of the validation. Thus, moving away from traditional waterfall approaches to a more agile approach is needed in the pharmaceutical industry.

Thus, XAI approaches should not solely be utilized after the development of an AI system to assure the reliability of an AI system. An early inclusion of XAI approaches could be beneficial for a more efficient development and validation process. However, the evaluation results highlight that this validation approach and also the use of an XAI component takes much more effort than traditional approaches.

Contribution to existing research The proposed validation approach and the results of its evaluation showed that XAI does help to collect sufficient evidence about the underlying decision behaviour of the AI to perform a software validation. Through the inclusion of steps like the evaluation with an XAI component and the specifying and testing of the boundaries of the AI sufficient evidence can be gathered to validate the AI. Thus, we demonstrated a way to reveal the inner workings of AI for a successful validation, which is one of the main challenges for the integration of AI systems in sensitive use cases. The evaluation further shows that it is better to incorporate important validation requirements already in the development process. Overall, it became clear that linear development approaches are less suitable. A more agile and iterative approach is needed allowing to step-wise discover the requirements and the underlying mechanisms of the AI. Both documentation and testing of AI based systems are different compared to traditional software. We conclude that AI validation objectives need to be considered early on and that the overall validation process is more laborious compared to traditional software validation due to the iterative nature. One major reason for increased efforts are insecurities caused by missing guidelines and regulations. In particular, guidelines are needed to support a more flexible, iterative approach in contrast to waterfall approaches favoured in current guidelines and proposals on how to deal with test and training data as well as how to perform AI testing.

6. Conclusions and Outlook

We propose XAI to validate AI systems in sensitive use cases. In a joint research and development project academic researchers, industry, technology, and GMP auditing experts developed a validation approach that was successfully applied in pharmaceutical manufacturing. However, aspects of the developed validation approach remain a case of assessment for the company implementing the AI systems. The validation approach is primary designed for sensitive use cases and might be too laborious for less sensitive use cases.Thus, we encourage future research to focus on the elaboration and establishing of standards and guidelines for sensitive use cases like the pharmaceutical industry. This may help to scale down the validation efforts needed and in consequence make thorough validation approaches more accessible also for less sensitive use cases.

References

[1] F. Königstorfer and S. Thalmann, "Applications of Artificial Intelligence in commercial banks–A research agenda for behavioral finance," *Journal of Behavioral and Experimental Finance*, vol. 27, p. 100352, 2020.

[2] J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the Need for Explainable Artificial Intelligence (xAI)," *arXiv preprint arXiv:2012.01007*, 2020.

[3] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020.

[4] P. M. Winter, S. Eder, J. Weissenböck, C. Schwald, T. Doms, T. Vogt, S. Hochreiter, and B. Nessler, "Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications," *arXiv preprint arXiv:2103.16910*, 2021.

[5] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Springer Nature, 2019.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[7] C. Meske and E. Bunde, "Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support," in *International Conference on Human-Computer Interaction*, pp. 54–69, Springer, 2020.

[8] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR, 2018.

[9] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.

[10] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities," *Information Systems Management*, pp. 1–11, 2020.

[11] M. Borg, "The AIQ meta-testbed: pragmatically bridging academic AI testing and industrial Q needs," in *International Conference on Software Quality*, pp. 66–77, Springer, 2021.

[12] J. Ochmann, S. Zilker, and S. Laumer, "The evaluation of the black box problem for AI-based recommendations: An interview-based study," 2021.

[13] U.S. Food and Drug Administration, "Code of Federal Regulations Title 21 Part 11."

[14] European Commission Health and Consumers Directorate-General, "EudraLex - Volume 4 - Good Manufacturing Practice (GMP) guidelines."

[15] ISO Central Secretary, "Quality management systems — Fundamentals and vocabulary," Standard ISO/TC 176/SC 1, International Organization for Standardization, Geneva, CH, 2015.

[16] U. S. Food and Drug Administration, "General Principles Of Software Validation; Final Guidance for Industry and FDA Staff," 2002.

[17] European Commission Health and Consumers Directorate-General, "EudraLex - Volume 4 - Good Manufacturing Practice (GMP) guidelines, Annex 11 - Computerized Systems," 2011.

[18] M. Schönberger and T. Vasiljeva, "Towards Computer System Validation: An overview and Evaluation of Existing Procedures," 2018.

[19] Amodei, Dario and Olah, Chris and Steinhardt, Jacob and Christiano, Paul and Schulman, John and Mané, Dan, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[20] M. Felderer and R. Ramler, "Quality Assurance for AI-based Systems: Overview and Challenges," *arXiv preprint arXiv:2102.05351*, 2021.

[21] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, 2020.

[22] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[25] S. M. Lundberg, G. G. Erion, and S. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," *CoRR*, vol. abs/1802.03888, 2018.

[26] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[27] International Society for Pharmaceutical Engineering (ISPE), "GAMP 5: A risk-based approach to compliant GxP Computerized Systems," 2008.

[28] U. S. Food and Drug Administration, "Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan," 2021.

[29] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, June 2005.

[31] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.