Aug 10th, 12:00 AM

Caution or Trust in AI? How to design XAI in sensitive Use Cases?

Anika Kloker
University of Graz, klokeranika@gmail.com

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

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

Thomas Kloiber
Leftshift One Software GmbH, thomas.kloiber@leftshift.one

Patrick Ratheiser
Leftshift One Software GmbH, patrick.ratheiser@leftshift.one

See next page for additional authors

Presenter Information
Anika Kloker, Jürgen Fleiß, Christoph Koeth, Thomas Kloiber, Patrick Ratheiser, and Stefan Thalmann

Caution or Trust in AI? How to Design XAI in Sensitive Use Cases?

Completed Research

Anika Kloker
University of Graz
klokeranika@gmail.com

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

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

Thomas Kloiber
Leftshift One Software GmbH
thomas.kloiber@leftshift.one

Patrick Ratheiser
Leftshift One Software GmbH
patrick.ratheiser@leftshift.one

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

Abstract

Artificial Intelligence (AI) becomes increasingly common, but adoption in sensitive use cases lacks due to the black-box character of AI hindering auditing and trust-building. Explainable AI (XAI) promises to make AI transparent, allowing for auditing and increasing user trust. However, in sensitive use cases maximizing trust is not the goal, rather to balance caution and trust to find the level of appropriate trust. Studies on user perception of XAI in professional contexts and especially for sensitive use cases are scarce. We present the results of a case study involving domain-experts as users of a prototype XAI-based IS for decision support in the quality assurance in pharmaceutical manufacturing. We find that for this sensitive use case, simply delivering an explanation falls short if it does not match the beliefs of experts on what information is critical for a certain decision to be reached. Unsuitable explanations override all other quality criteria. Suitable explanations can, together with other quality criteria, lead to a suitable balance of trust and caution in the system. Based on our case study we discuss design options in this regard.

Keywords

Artificial Intelligence, XAI, Trust, Case Study, Pharmaceutical Industry

Introduction

Artificial intelligence (AI) based on machine learning (ML) has growing business potential and is applied in most industries. However, it is still lagging in sensitive use cases, e.g., those with a high-risk classification in the proposed EU AI regulatory framework on artificial intelligence. Especially regulated industries, such as finance or the pharmaceutical industry, are not yet able to make full use of AI (Königstorfer and Thalmann 2021; Polzer et al. 2022). When AI decisions affect human lives, sensitive data, or may cause serious risks, AI needs to be trustworthy and reliable (Xu et al. 2019).

Trust and auditing problems, at least partially, arise from the AI's black-box character (Carabantes 2020), which drew more attention with the increase of potential application areas (Gerlings et al. 2021). Black-box AI is often opaque and cannot be understood by users (Guidotti et al. 2019), thus it suffers from a lack of acceptance and trust (Carabantes 2020; Ribeiro et al. 2016). Research on explainable AI (XAI) aims to mitigate the black-box by developing approaches to make both the decision-making principles of the AI as a whole (global explainability) and the reasons for individual decisions (local explainability) transparent. XAI is defined as AI system, which explains its reasoning and helps to characterize the weakness and strengths of the system. Also, information to help expect future behavior of the system is conveyed to a

human user (Gunning and Aha 2019; Meske et al. 2020). Such transparency can then lead to both increased trust by users and the ability to audit the system (Adadi and Berrada 2018).

Trust and acceptance are necessary for the successful implementation of an IS (Venkatesh et al. 2016), especially for AI (Wang and Siau 2018). So far, literature intends XAI to maximize trust in AI-based IS, predominantly from a consumer perspective. However, in an industrial context and especially for sensitive use cases, not a maximum of trust but rather appropriate trust as a result of balancing caution and trust is desired. Blind trust would undermine the requirement of human oversight for sensitive use cases. Blind trust can and should be prevented by XAI (Jacovi et al. 2021). Explanations reveal the capabilities of the system and therefore allow users to find an appropriate level of trust (Lee and See 2004; Weitz 2021). So far, research on how to design XAI in AI-based IS to balance caution and trust is missing. We thus address the following research question: What are design-factors of XAI that can help achieve a suitable balance of caution and trust in sensitive use cases?

To the best of our knowledge, we conduct the first study on XAI and its effects on achieving a suitable balance of caution and trust in a use case of a regulated industry. For this purpose, a team of industry experts and academics conducted a case study about deploying an AI-based prototype in pharmaceutical manufacturing. The prototype is used in a routine production environment to evaluate AI-supported documentation and classification of quality events.

Background

As the complexity of AI increases, e.g., with deep learning (DL), this leads to an increase in their performance while at the same time the opaqueness increases (Samek and Müller 2019). Thus, the mechanisms generating an output based on input data become hidden inside a black-box (Adadi and Berrada 2018). For such opaque AI, users are unable to understand how the decisions are made by the AI (Xu et al. 2019). However, explanations of how the AI works and how individual decisions were made, are important factors to foster users' trust in the AI (Barredo Arrieta et al. 2020; Ribeiro et al. 2016).

XAI approaches have been developed to explain how an AI makes decisions varying in providing explanations. The XAI approach has a respective impact on the different stakeholders and their trust relationship towards AI (Rai 2020). In general, XAI approaches are either *local* or *global*. The former aim to enable an understanding of the model as a whole, while the latter provide insights into why a specific decision was made (Danilevsky et al. 2020). Additionally, not every AI needs an additional approach to make it understandable, some types of models are transparent themselves (e.g., decision trees) while more complex models (e.g., DL) require post-hoc explanations (Barredo Arrieta et al. 2020).

The use case studied in the paper at hand uses an AI-based system in the field of Natural Language Processing (NLP). For NLP, specific XAI approaches based on visualizations have been developed (Danilevsky et al. 2020). To explain, e.g., the XAI technique *Feature Importance*, visualization methods like *Saliency Heatmapping* or *Saliency Highlighting* have been adapted (Danilevsky et al. 2020). The *Feature Importance* approach is about representing the importance of the features that led to the output of the prediction (Danilevsky et al. 2020).

XAI can foster the successful integration of AI in practice by building a trustworthy AI (Weitz 2021). This is a necessary precondition, especially in an industry context where (a) the trust of professional users in the AI must be established, as trust leads to acceptance and (intention to) use (Wang and Siau 2018). In addition, (b), IS used in industry need to be made reliable and accountable (Ryan 2020). This is especially important in sensitive areas (Xu et al. 2019), even though there often are requirements to have a human in the loop making the final decision (Stuurman and Lachaud 2022). Thus, in such cases, decision-making by experts should be supplemented, but not replaced by it (Sutton et al. 2018). However, the acceptance of AI for decision support also depends on trust in the system (Wang and Siau 2018).

In sensitive use cases, it is important that trust also remains limited, as blind trust in the system could lead to overestimating its capabilities (Parasuraman and Manzey 2010) or overreliance on the system leading to increased risk-taking (Wagner et al. 2018). Blind trust or too much trust in AI can lead to severe negative consequences for the affected users (Goddard et al. 2011). Thus, for sensitive use cases, appropriate trust is the goal that is high enough to encourage use within the restrictions of the system's capabilities. This can be achieved by revealing the true capabilities of the decision-making system to the users (Lee and See 2004)

and to facilitate a correct mental model of the system's workings (Schraagen et al. 2020). Thus, XAI may create distrust in a non-trustworthy AI-System (Jacovi et al. 2021). Trust and distrust may coexist, while the aim, in this case, is the creation of general trust in AI, but keep the user cautious and verify each decision (Lewicki et al. 1998). XAI may also point out potential pitfalls and problems that might not be detectable using traditional metrics (Polzer et al. 2022). How XAI affects this relationship of caution and trust toward AI and how to design XAI for appropriate trust has not yet been investigated. This research gap will be addressed by this paper.

Case Study

The case study was conducted together with a pharmaceutical manufacturer with around 1600 employees in Austria. The manufacture of sterile pharmaceuticals (injectables) requires stringent adherence to detailed specification documents and compliance with validated processes. Highest product quality is of utmost importance, therefore, personnel carrying out these manufacturing processes are specially selected and intensively trained. Processes are also continuously monitored but as in any industrial process, however, fluctuations happen and more or less serious incidents occur. In such a case, the pharmaceutical quality assurance system requires timely, precise documentation of these incidents and an evaluation process ("quality event process") (EC 2017; "US FDA" 2022). As a rule, the documentation must be carried out by the person who first discovers the defect or incident, e.g., machine operators. In event processing, possible effects on the end product (and thus on the patient) must be thoroughly evaluated. Depending on the criticality of the incident, a distinction can be made between (non-critical) "incidents" and (critical) "deviations". The depth of analysis and consideration may naturally be less in the former case. Finally, as part of pharmaceutical checks and balances, only members of Quality Assurance (QA) can conclude the event. From a business perspective, primarily two reasons drive up required resources tied to the process: (1) initial documentation quality and (2) criticality of the event.

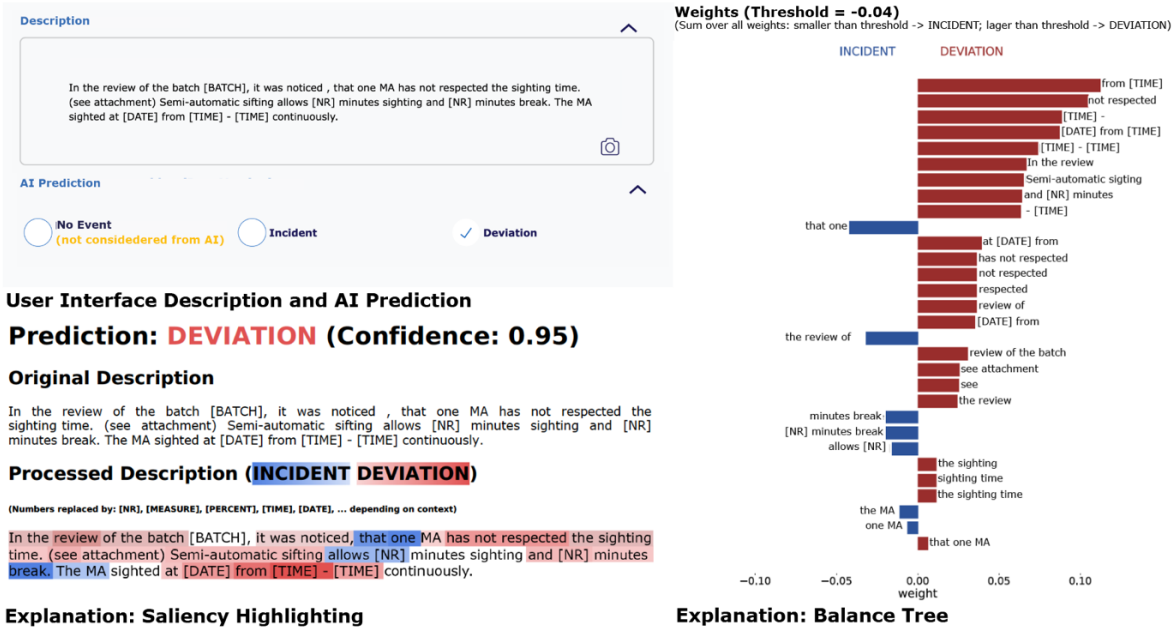

Figure 1: Prototype XAI

The AI prototype is intended to address these cost drivers by supporting employees in the initial event documentation and by providing a real-time classification of events. However, due to the regulatory requirements, the AI should only provide decision support and the final decision needs to be taken by the employee. In this regard blind trust and overreliance is undesired.

AI Prototype to predict Event Types from unstructured Descriptions

The AI prototype aimed to classify the severity of an event as either "Incident" or "Deviation" based on a short description, i.e., a binary, NLP classification task. The available data consisted of approx. 15.000 events, each consisting of 50 features. For model development, only a single feature ("Description") could be used for training and inference and the label was given by the feature "Event Type". Model selection and hyper-parameter tuning were based on the averaged f1-score on the validation set. The overall best performance with an f1-score of 0.85 on the test data was obtained with a classical statistical approach to text classification. Besides having achieved the highest score, the logistic regression model had also one of the simplest structures among the studied models and therefore is especially amenable to explainability techniques, however, the interpretability of the generated explanations was no factor in model selection.

Following the classification scheme of Danilevsky et al. (2020), a local, self-explaining approach was adopted to explain the model predictions for individual texts. Due to the nature of the model, every n-gram (word/phrase) of an input sample is associated with a numeric weight (combination of tf-idf and regression weights). This can be shown in tabular form for each inference, allowing – in principle – manual tracing of the model's prediction steps from the raw text input. Two types of visualizations for these explanations were generated (see Figure 1): 1. a bar-plot style overview showing a balance of the most important phrases in the text for both classes and their respective weights, and 2. a two-colour saliency highlighting of the whole description text, where each colour corresponds to a class and different shades indicating the importance (weight). The confidence of the AI's decision-making is also provided.

Procedure

The study aims to identify what factors influence trust or caution of users. From the development through the test implementation in the company, the project team conducted regular workshops. The requirements for the AI were elicited and it was determined which characteristics a XAI visualization must fulfill. Through a literature review, the variety of NLP-AI methods, their corresponding explanation and visualization capabilities were explored, and two XAI-visualization methods were selected and implemented (see Figure 1) (Danilevsky et al. 2020). An interview study in two sessions was conducted with seven employees of the quality assurance process from the pharmaceutical company. For the first session, a total of 15 real cases were selected from a pool of 40 cases for the interviewees to evaluate. Attention was paid to varying prediction confidence levels and also to the inclusion of erroneous AI decisions. In the second session, a new pool of 40 cases was provided by the pharmaceutical company. To identify the effect of XAI on caution and trust, interviewees used the prototype under three conditions to assess 15 cases in each condition, again with a prediction of varying confidence levels and also incorrect decisions. Each interviewee completed all three conditions in the order displayed in Figure 2.

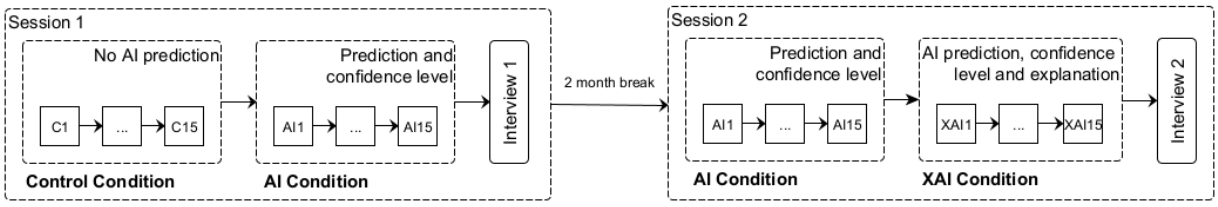

Figure 2: Conditions and Interview Sequence

To separate the condition with and without XAI, interviews were conducted in two separate sessions with a two-month break in between. In each session, interviewees first used the prototype, followed by an interview. In the first session, real case assessments were conducted in the Control and AI Conditions. Interviewees were encouraged to constantly verbalize their thoughts when assessing the cases following the advice of the think-aloud method (Jaspers et al. 2004). The cases in the Control and AI Conditions were the same. Interviewees were asked for their assessment of the case and how they perceive the AI and its performance. In the second session, interviewees were first allowed to assess the real cases on their own in the AI condition and subsequently were provided with the same cases in the XAI condition with the visualizations of the explanations. As in session 1, the subsequent interview aimed to determine the changes in perception regarding the AI and the confidence in it and its performance. The prototype frontend was

replicated using oTree (Chen et al. 2016) ensuring that changes to the production prototype between interviews could not lead to different AI recommendations and explanations.

Interviews were recorded, transcribed and then analyzed according to the recommendations of thematical analysis (Braun and Clarke 2006). In the first coding of the transcript, initial codes were set. Within these initial codes patterns of themes could be developed. The overarching themes were split up into three fields: Irritation/Distrust/Trust in AI, Understanding AI through XAI and Irritation through XAI. As a final coding step, the exact barriers to trust in AI were highlighted.

Results

The comparison of the interviews from sessions 1 and 2, respectively, allowed us to identify factors influencing caution and trust regarding AI depending on the explanations provided. A total of five factors were derived: Suitability of Explanation, Accuracy, Process Integration, System Knowledge and Confidence Level AI will be presented with consideration of the influence from the application of XAI.

Suitability of Explanation

We study the question of how explanations impact users' caution or trust in AI. To be able to isolate the effect of XAI, AI decisions without XAI were presented in session 1. However, even in the first session, the need for explanations of how decisions are reached was communicated: *"I think you need the decision-making basis of the AI to be able to understand that somehow. Why? Why is it decided this way*? [Interview Partner (IP)3]". The interviewee highlights that she is unable to judge the AI without more detailed information about the decision. This uncertainty about how the system works fosters caution.

While a missing explanation facilitates caution, a suitable explanation can create trust as one interviewee answered the question if explainability increases his trust level in the AI: "*If the explanation fits my assessment, then clearly yes. But if the AI does a rating on the basis of some words I cannot comprehend, then not*." [IP3]. The interviewee perceived the explanation as decision basis and this explanation needs to fit into the established reasoning of the user or at least should make sense for him to increase trust. Other interviewees state even more clearly that unsuitable explanations lead to caution. On the question of how explanations affect the interviewees' trust in the AI, another interviewee answered: "*Definitely in my case negative. […] When I look at these evaluations, you would have thought, it would have used a little bit more meaningful criteria to assess that*." [IP5]. The interviewee complains that the criteria used for the explanation are not meaningful to him. As a consequence, the interviewee not only has doubts about how the systems works, but he also assumed the system's decision is based on something which does not relate to the decision problem, even though the predictive accuracy of the AI system was very high. Several interviewees complained about the provided explanations demanding that they need to be suitable. Explanations are suitable if the explanation fits the established reasoning of the users. Participants also work with keywords in their assessment process itself. If these keywords are wrongly categorized or wrongly weighted by the AI, caution is triggered.

Accuracy

A factor influencing caution or trust in the AI mentioned by our interviewees is the accuracy of the AI. Interviewees are highly sensitive regarding wrong classifications and any obvious wrong classification or doubts in recommendations increases caution. In contrast, interviewees are less cautious if the AI works well as one interviewee said: *"I believe I would be less cautious if I know the AI has an accuracy of 99%."* [IP1]. The interviewee wants to be convinced of the high accuracy and that in this case, our interviewees reported that this leads to trust in the AI.

After the application of XAI, this logic is changed significantly, and the effect of accuracy was overshadowed by the suitability of the explanation: "*As I said, this is really interesting for me that with those criteria, which are not the reason for the classification from my perspective, it works so well. […]I am really interested why this works, but as it works as it works, I would never use it*." [IP5]. Thus, even if the system has a (for him surprisingly) high accuracy, he is cautious and would not use the system if the explanation is unsuitable. But our interviewees reported that a suitable explanation combined with a high accuracy clearly increases the trust in the AI. In this regard, we discussed the tradeoff between accuracy and suitability of

explanations with our developer team and the pharmaceutical company. We concluded that in case no suitable explanations are provided, the accuracy is the key criterion and the interpretability of the features (which determine the suitability of the explainability) is less important. However, in case a suitable explanation is offered, the accuracy may need to be relaxed to increase the interpretability of the features and thus, the suitability of the explanations.

Process Integration

Another factor determining trust or caution in AI is the process integration of the AI. This is the scope the AI considered in relation to the entire process, which is the subject of the AI assessment. If the AI focuses only on selected sub-problems of the business process and thus neglects its complexity, the participants state that the AI is lacking an understanding of the technical procedures and their background*: "I cannot trust the system as the context is partly considered wrong or not at all, further there are further aspects which are currently not considered at all."* [IP3]. In addition to the limited consideration of context for the system interactions he had so far, the interviewee also complains about the limited flexibility for new situations. On the one hand, the limited process integration and limited consideration of the context become directly apparent in the AI's incorrect predictions. On the other hand, the participants also assume this from the technical design of the AI, since they know certain procedures that are not included in the system design and the training data. One interviewee stated: *"[T]he things that are still behind it, […] can lead to a different evaluation, which in my opinion cannot be covered by the AI*." [IP2].

The assumed limited scope of the AI was frequently mentioned and subsequently confirmed by inspecting the explanations of the XAI. One interviewee explained his surprise about the explanation: *"[I]n this situation, I need to consider <X> and everything else (explanation provided) is irrelevant. This is the reason why I am surprised that the AI made the decision most of the time always correct. […] What I see is why this is not expedient"*. [IP5]. The employee expects <X>, but as this is outside the scope of the system it is not part of the explanation. Consequently, he remains skeptical even if the system worked precisely so far. In the previous round, it was mentioned several times that it is necessary to include more information through consultation or basic process knowledge, in a decision. By presenting the keywords or keyword-terms and their corresponding weights for the decision, the participants notice that "*the keywords are valued incorrectly or in the wrong context"* [IP2]. Even if the AI's decision turns out to be correct, it is apparent that due to a lack of process integration, the AI decides with a missing or incorrect context. This leads to caution among the interviewed experts.

System Knowledge

Knowledge about the AI-based IS, composed of the experience and information the user has about the system, is another important factor. This includes information the user has about the development and the AI itself. Who programmed the AI and, above all, on which dataset is the model built? If a good reputation can be assumed here and the users know and trust this basis, they can also trust the predictions of the AI. Otherwise, interviewees become cautious: *"For me, it is challenging to judge the quality of the system as I don't know about the underlying datasets. […] I don't believe that the system works 100% and even if the AI gets more and more data, the quality of the labels is not always so good that an AI can work with it."* [IP3]. The interviewee highlights that the training data sets are often not perfect and include inconsistencies. Thus, if even humans cannot reach a clear consensus, how can an AI learn from this?

Experience with the system also contributes to system knowledge building on experience of prediction accuracy as well as the limitations of the system. One participant describes the trust-building process: *"If I see that the AI is now making the correct decision over 3 years or whatever, that if that trust has really established itself over the long term, then maybe I would also move to say okay, now I'm going to listen to that."* [IP5]. Lack of experience over time is thus reducing trust and raising caution.

The explanation shows the user the factors and the corresponding weights the AI decision is based on. If the explanation does not fit (see suitable explanation) or the explanation reveals shortcomings of the system (lack of process integration) the users partly refuse to use the system. Thus, users very quickly form negative opinions as this explanation *"clearly shows how the system works"* [IP7]. Hence, explanations need to be used with care, because as one interviewee said: *"This is a learning journey and decisions become better*

and better the longer I use the system" [IP3]. Thus, it needs to be ensured that users engage with the AI for a certain period of time to build critical system knowledge.

Confidence Level AI Decision

A correct decision of the AI with a high confidence level inspires trust in the system. Displaying the AI's confidence level in its decision was particularly crucial to an impression of the AI in the first session without XAI. If the confidence level deviates strongly from the confidence that the participants themselves have in the decision, this leads to irritation: *„No, for me it is described very clearly. So I'm surprised that it's only 88 percent."* [IP5]. Likewise, a low confidence level of the AI triggers caution toward AI predictions. For example, one participant notices, that the AI cannot make a reliable decision at a confidence level of 4%. *„For the initial assessment, it's a quick assessment and a direction, because if you look at the percentages, you can only get in the 4 percent and so on. So, the AI doesn't really know much either."* [IP7]. By switching on the explanation visualization in the second session, the display of the AI's confidence level lost attention. In addition, now cases were predicted correctly, with a high degree of certainty, but the XAI displayed weighted keywords that did not fit the participants' reasoning: *"It should have been the other way around. The evaluation. So, it's exactly the other way around."* [IP4]. The caution caused by a low confidence level and its visualization among the participants give rise to another challenge. The participants recognize through the visualization of the explanation the narrow scope of the AI training data set. Thus, one participant also describes that the textual contents of the case, are not sufficient for a proper assessment and decision: *"I would let the AI decide more strictly in the case of doubt. Because from my point of view, it's always better in the worst case if there really is this uncertainty and the description doesn't give me what I need to be able to classify."* [IP5]. The threshold for the AI to decide at all should therefore be set higher. A system that makes decisions based on inadequate data, which becomes visible through the XAI, is not seen as trustworthy.

Designing a valuable and trustworthy XAI – Discussion

Our results show that the successful use of XAI depends on several different criteria which we outline below. We make two major contributions: We contribute to the (1) understanding of the suitability of XAI by investigating regulated use cases and (2) identify factors relevant to designing XAI in regulated use cases to achieve an appropriate balance of trust and caution.

Suitability of XAI

The suitability of both the type of explanation and the explanation about the AI itself is the dominant factor influencing users' caution or trust. Trust in an AI can be reduced by explanations revealing a decision logic of the system not fitting to the users' mental model. If an unsuitable explanation prevents users' trust and acceptance of the system, the system's successful implementation is at risk (Sovrano et al. 2021). Thus, first (1) a model must be chosen that can be explained, either by itself as a transparent model or a post-hoc explanation. Second, (2), the objective of the explanatory content must be determined. Should the general logic of the system or/and the reasoning of individual decisions be explained? Should information about the performance or input data of the AI be provided (Liao and Varshney 2022)? Then, (3), it must be possible to present this explanation using a suitable visualization. One should keep in mind that making something explainable is not the same as explaining it (Sovrano et al. 2021). However, when aiming to implement a trustworthy XAI, simply providing explanations or interpretable models is not enough on its own. Thus, (4), the content of this explanation must be acceptable to the users of the system. In this feedback process, it can then also be determined whether the explained AI follows decision paths that fit the expectation on how and why a decision should have been made of the user. If the model is fundamentally unable to fit the expected decision path and reasoning of the user, it will be treated with an overabundance of caution. But users might develop too much trust in a system if no explanations are given for correct decisions as potentially a wrong decision logic could not be discovered and lead to wrong decisions in future cases (Polzer et al. 2022).

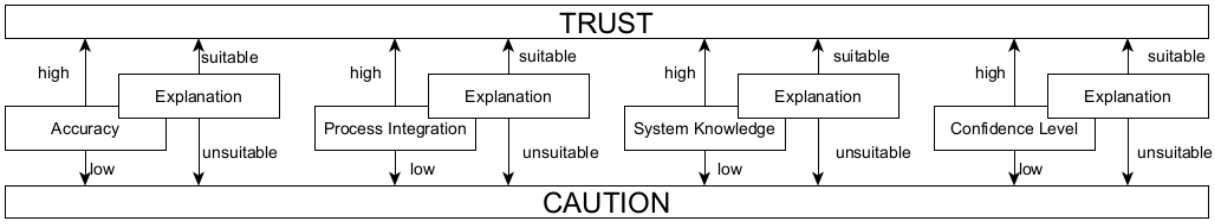

Figure 3: Factors influencing the Trust and Caution Balance

The disproportionality large impact of explainability becomes clear in Figure 3. If the explanation is unsuitable, an AI may be quickly rejected. Even if the explainability, by theory, of a system can be fulfilled by, e.g., a transparent model (e.g., regression like in the case at hand) and the other factors for an appropriate trust-relationship are high, the trust and acceptance in the system can nevertheless be negated if the explanation is not suitable.

Designing XAI-based IS in Sensitive Use Cases

The suitability of the XAI component is dominating in a professional context, However, the factors accuracy, process integration, system knowledge and confidence level impact the trust and caution balance. They can shift the balance in the direction of caution if they are low. If they are high, they can increase trust, but only if the suitability of XAI is satisfied.

The overall **accuracy** of the system is of course important when choosing a model, as higher accuracy leads to higher trust, especially as results from the algorithm aversion literature show that errors made by algorithms disproportionally decrease trust as compared to a human (Dietvorst et al. 2015). In addition, options to explain the chosen model need to be considered and balanced. Usually, this is discussed as a trade-off between accuracy and explainability (Barredo Arrieta et al. 2020). In our case, however, it is especially important to consider that we found that the positive effect of a high accuracy only applies if the explanation is suitable for the user. Thus, when designing an AI for domain experts as users in a high-risk use case, development has to be an iterative process of model training, explaining and assessment of the explanation by the domain experts. A focus on accuracy as a key measure alone is not enough. If the explanation is not deemed suitable by domain experts, another model should be applied, the performance of which may not be as high, but which may be deemed trustworthy.

For **process integration**, the better the AI is integrated and demonstrates this in the explanations of its decisions, the more those decisions will be trusted. However, perfect integration might not be desirable as it is difficult to adequately represent all processes. However, increased process integration can increase users' perception that the uniqueness of the different situations is taken into account thereby increase willingness to use the system (Longoni et al. 2019). However, perfect integration might not be desirable as it is difficult to adequately represent all processes with the respective depth, especially with a well-explainable model. In addition, this may shift the trust and caution ratio too far in the direction of trust. From the perspective of appropriate trust this is not desirable in a high-risk context (Grigsby 2018), thus focusing on one part of the whole process seems suitable and maintains caution about the decision. However, this in turn might reduce trust too much, as users assume that the system will not be able to solve future problems, especially if contexts not included in the information the AI uses change.

Knowledge about the system determines the correctness of the mental model users have about the AI-based IS. Note that even without communication and explanations about the system, users form a mental model of it. In our case, it was based on the first cases and AI recommendations that the user encounters. This can lead to an inappropriately high level of trust or caution (Schraagen et al. 2020). A suitable XAI will communicate the reasoning and allow, together with information about the system, used datasets, accuracy etc., a user to form an appropriate mental model of the system and through this, a level of trust appropriate to the capabilities of the system. If this can be successfully achieved, however, is dependent on the objectivity of the task itself; as our results showed the classification of events is not always straight forward and such a lack of objectivity is seen as difficult to solve by AI in the perception of users (Castelo et al. 2019).

Finally, the **confidence level** given to the user for decisions has a strong impact on the user's level of trust or caution. A high confidence level is more likely to elicit trust from the user, while a low level for even a

single decision may raise caution (Zhang et al. 2020). Making and communicating decisions with a low confidence level should therefore be avoided. In this case, the system should not make a decision in a confidence level range that triggers unwanted levels of caution. This is compounded by the fact that explanations of individual decisions with a low confidence level will more likely be inappropriate and let the user create a mental model underestimating the system's capabilities.

Finally, our work is not without **limitations**. We studied one AI-prototype for one specific setting and thus do not claim generalization of the results. Future work can investigate our explorative work in more detail. To reflect any poor data input quality, the prototype should also have the output option of coming to no result. Further, the confidence level scaling was not specified and thus decisions with low confidence were presented to the users. This is less likely in practice but also offered us the opportunity to study this phenomenon. The interviews were conducted in German and quotes were translated for this paper. Two authors double-checked the translation to avoid translation issues.

Conclusion & Future Work

We presented the first study on design factors to influence trust and caution in AI in pharmaceutical manufacturing. Based on a prototype and interviews with employees of a real-world quality assurance process we showed that XAI, together with other more traditional factors, can not only increase trust but rather lead to caution. Thus, when designing AI in sensitive use cases with domain experts as users, an iterative process between model training, explaining and evaluation by domain experts must be employed. This can result in reaching appropriate trust so that users' trust allows its successful implementation while ensuring that users remain appropriately cautious to maintain meaningful human oversight. In the future, these findings should still be applied and researched on more complex AI systems and in more dynamic environments. Accordingly, different effects on the trust-caution balances are to be expected. Likewise, the elaborated factors should be investigated in use cases with expert users and in different industries.

REFERENCES

Adadi, A., and Berrada, M. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* (6), pp. 52138–52160.

Barredo Arrieta, A. et al. 2020. "XAI: Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115.

Braun, V., and Clarke, V. 2006. "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology* (3:2), pp. 77–101.

Carabantes, M. 2020. "Black-Box Artificial Intelligence: An Epistemological and Critical Analysis," *AI & SOCIETY* (35:2), pp. 309–317.

Castelo, N., Bos, M. W., and Lehmann, D. R. 2019. "Task-Dependent Algorithm Aversion," *Journal of Marketing Research* (56:5), pp. 809–825.

Chen, D., Schonger, M., and Wickens, C. 2016. "OTree—An Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance* (9), pp. 88–97.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. 2020. "A Survey of the State of Explainable AI for Natural Language Processing," *ArXiv:2010.00711 [Cs]*.

Dietvorst, B., Simmons, J., and Massey, C. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err.," *Journal of Experimental Psychology* (144:1), pp. 114–126.

EC. 2017. *EudraLex The Rules Governing Medicinal Products in the European Union Volume 4 Good Manufacturing Practice*.

Gerlings, J., Shollo, A., and Constantiou, I. 2021. "Reviewing the Need for Explainable Artificial Intelligence (XAI)," *ArXiv:2012.01007 [Cs]*. (http://arxiv.org/abs/2012.01007).

Goddard, K., Roudsari, A., and Wyatt, J. C. 2011. "Automation Bias - a Hidden Issue for Clinical Decision Support System Use," *Studies in Health Technology and Informatics* (164), pp. 17–22.

Grigsby, S. S. 2018. "Artificial Intelligence for Advanced Human-Machine Symbiosis," in *Augmented Cognition: Intelligent Technologies* (Vol. 10915), Springer LNCS, pp. 255–266.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2019. "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* (51:5), pp. 1–42.

Gunning, D., and Aha, D. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine* (40:2), pp. 44–58.
Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. 2021. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI," (http://arxiv.org/abs/2010.07487).
Jaspers, M., Steen, T., Bos, C., and Geenen, M. 2004. "The Think Aloud Method: A Guide to User Interface Design," *International Journal of Medical Informatics* (73:11–12), pp. 781–795.
Königstorfer, F., and Thalmann, S. 2021. "Software Documentation Is Not Enough! Requirements for the Documentation of AI," *Digital Policy, Regulation and Governance* (23:5), pp. 475–488.
Lee, J. D., and See, K. A. 2004. "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society* (46:1), pp. 50–80.
Lewicki, R. J., McAllister, D. J., and Bies, R. J. 1998. "Trust and Distrust: New Relationships and Realities," *The Academy of Management Review* (23:3), p. 438.
Liao, Q. V., and Varshney, K. R. 2022. "Human-Centered Explainable AI (XAI): From Algorithms to User Experiences," (http://arxiv.org/abs/2110.10790).
Longoni, C., Bonezzi, A., and Morewedge, C. K. 2019. "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research* (46:4), pp. 629–650.
Parasuraman, R., and Manzey, D. H. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors* (52:3), pp. 381–410.
Polzer, A., Fleiß, J., Ebner, T., Kainz, P., Koeth, C., and Thalmann, S. 2022. *Validation of AI-Based Information Systems for Sensitive Use Cases: Using an XAI Approach in Pharmaceutical Engineering*, Proceedings of HICSS.
Rai, A. 2020. "Explainable AI: From Black Box to Glass Box," *Journal of the Academy of Marketing Science* (48:1), pp. 137–141.
Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," (http://arxiv.org/abs/1602.04938).
Russell, S. J., and Norvig, P. 2021. *Artificial Intelligence: A Modern Approach*, (Fourth edition.), Pearson Series in Artificial Intelligence, Hoboken: Pearson.
Ryan, M. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability," *Science and Engineering Ethics* (26:5), pp. 2749–2767. (https://doi.org/10.1007/s11948-020-00228-y).
Samek, W., and Müller, K.-R. 2019. "Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700), Springer LNCS pp. 5–22.
Schraagen, J. M., Elsasser, P., Fricke, H., Hof, M., and Ragalmuto, F. 2020. "Trusting the X in XAI: Effects of Different Types of Explanations by a Self-Driving Car on Trust, Explanation Satisfaction and Mental Models," *Proc. of the Human Factors and Ergonomics Society*, pp. 339–343.
Sovrano, F., Vitali, F., and Palmirani, M. 2021. "Making Things Explainable vs Explaining: Requirements and Challenges under the GDPR," *ArXiv:2110.00758 [Cs]* (13048), pp. 169–182.
Stuurman, K., and Lachaud, E. 2022. "Regulating AI. A Label to Complete the Proposed Act on Artificial Intelligence," *Computer Law & Security Review* (44), p. 105657.
Sutton, S., Arnold, V., & Holt, M. 2018. "How Much Automation Is Too Much? Keeping the Human Relevant in Knowledge Work," *J. of Emerging Technologies in Accounting* (15:2), pp. 15–25.
"US FDA." 2022. *21 CFR Chapter 1, Subchapter C, Part 211, "Written Procedures, Deviations."*
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. "Attention Is All You Need," (http://arxiv.org/abs/1706.03762).
Venkatesh, V., Thong, J., Xu, X. 2016. "Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead," *JAIS* (17:5), pp. 328–376.
Wagner, A. R., Borenstein, J., and Howard, A. 2018. "Overtrust in the Robotic Age," *Communications of the ACM* (61:9), pp. 22–24.
Wang, W., and Siau, K. 2018. "Building Trust in Artificial Intelligence, Machine Learning, and Robotics," *Cutter Business Technology Journal* (31(2)), pp. 47–53.
Weitz, K. 2021. "Towards Human-Centered AI: Psychological Concepts as Foundation for Empirical XAI Research," *It - Information Technology*.
Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. 2019. "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," in *Natural Language Processing and Chinese Computing* (Vol. 11839), Springer LNCS pp. 563–574.
Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. 2020. "Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305.