Study on Criteria for Explainable AI for Laypeople

Thorsten Zylowski[1,2]

[1]University of Hohenheim, Schloss Hohenheim 1, 70599 Stuttgart, Germany
[2]Karlsruhe University of Applied Sciences (HKA), Moltkestraße 30, 76133 Karlsruhe, Germany

Abstract

Artificial intelligence (AI) is increasingly influencing everyday situations. Humans are becoming more and more dependent on the decisions made by AI. Due to the black-box nature of AI models, the decisions of these models can hardly be understood by AI experts and certainly not by laypeople. This results in a potential trust problem in systems that use AI. Methods from the field of Explainable AI are being used to try to counteract these trust problems. Unfortunately, most of the methods are designed by AI experts for AI experts and do not take into account the specific requirements of laypeople. This paper presents a study on criteria for Explainable AI for laypeople to help design AI systems that meet the needs of laypeople and aim to increase trust. A survey was conducted with 103 participants, exploring different areas for the design of Explainable AI for laypeople. These include the importance of explanations, the extensiveness of an explanation, which interactive elements are useful, and the role of AI and human certainty when making decisions with explanations.

Keywords

User Study, Explainable AI, Laypeople, Trust

1. Introduction

The influence of artificial intelligence (AI) is increasing in more and more everyday situations. Examples include traffic routing, Internet searches, movie recommendations, as well as more critical areas such as insurance contracting and university admissions. Humans are becoming more and more dependent on the decisions of AI. Unfortunately, the decision paths and the workings of the machine learning (ML) models that are used can only be understood by experts, if at all. In the past, algorithms were developed by engineers, and the process, even if complicated, could be understood by a wide range of people, including laypeople in many situations. Today, these algorithms are learned by AI from an enormous amount of data. The resulting models consist of countless weighted connections that, even if a person have insight into these weights, he or she cannot understand. Consequently, the models can be considered a black box. From this characteristic arises the need for transparency, since without transparency, trust in the AI algorithms as well as in the results cannot be built. In order to gain more insight into the models and to better understand decision-making processes of AI and thus providing transparency, several methods have been developed. These can be used to examine the effect of different

Proceedings of the Second International Workshop on Explainable and Interpretable Machine Learning (XI-ML 2022) co-located with the 45rd German Conference on Artificial Intelligence (KI 2022), September 20, 2022, Trier (Virtual), Germany

thorsten.zylowski@uni-hohenheim.de (T. Zylowski)
0000-0002-5029-898X (T. Zylowski)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3457/paper3ximl.pdf

inputs to the model on outputs to understand what internal processes have been learned. The set of these methods is summarized by the term Explainable AI (XAI). There are methods designed for explanations of specific architectures (e.g. [1]), as well as model-agnostic approaches that can be applied to a variety of models (e.g. [2]). In addition, methods can be categorized into local methods that explain individual AI decisions or recommendations (e.g. counterfactuals [3]) and global methods that explain the model as a whole (e.g. model class reliance [4]). However, the field is still in its infancy, which means that the overwhelming majority of the methods have been created by AI experts for AI experts. However, if AI experts understand the system better, it doesn't mean that laypeople do as well. It follows that there is a great need for Explainable AI methods that address the requirements of laypeople, helping them to understand AI decision processes, increase transparency and ideally work as a basis for trust. In order to design Explainable AI methods that take into account requirements of laypeople, it is important to systematically collect these requirements. Therefore, a study was conducted to capture the essential criteria of Explainable AI for laypeople. A key focus was on examining trust-building aspects of the methods. With the results from this study, Explainable AI methods can be designed and tested in future work for their suitability for laypersons and hopefully help to increase trust in AI systems.

2. Related Work

A frequently cited work on what are good explanations for humans is [5]. He presents insights from many years of social science research for the use in Explainable AI. For him, a good explanation is, first, *contrastive*. Humans ask not only why an event occurred, but why it occurred rather than another event. Second, good explanations are *selected*. It is impossible to state the complete causal chain of events, because it is potentially infinite, since there is always a preceding cause. Humans select explanations from a set of explanations and take them as *the* explanation. Irrelevant and already known causes should not be provided. Third, he goes into further detail, that *"probabilities probably don't matter"*. This implies that people may choose the best explanation based not on probability but on other metrics, such as simplicity and relevance. Fourth, he states, that explanations are *social*. Explanations should be tailored to the person receiving them and take into account the social context. Often explanations are given in dialogues, iteratively providing more detailed information as needed. As [5] states: *"explainer and explainee may interact and argue about this explanation"*. [6] made an extensive literature review of Explainable AI and categorized it into the five groups (1) Definition of explanation, (2) Goals of explanations (WHY), (3) Content to include in the explanation (WHAT), (4) Types of explanations (HOW) and (5) Evaluation of explanations.

They *"critically review the previous sections and give insights on new directions to create better explanations"* [6]. They argue that AI systems should provide more than one explanation, targeting different groups of people: (1) Developers and AI researchers, (2) Domain experts and (3) Lay users.

They analyze how the needs for explainable system proposed by [7] can be brought together with these groups. While verification and improvement of the system through explainability are needs of the first two groups, laypersons are mainly concerned with the right to an explanation.

Especially in situations that can affect their lives, people want an AI system to explain its decision. In addition, they argue, like [5], that explanations should follow cooperative principles of human conversation. For laypeople, they specifically recommend, based on existing technical capabilities, to offer explanations with multiple counterfactuals from which people can interactively choose the appropriate explanation. They argue that *"this explanation is parallel to human modes and it is very likely to generate trust"*. Both [5] and [6] refer to work by [8] and argue that explanations should follow his maxims of good human conversation, i.e. *quality* (only say what you believe), *quantity* (only say what is required), *relation* (only say what is relevant) and *manner* (say it nicely).

In addition to these very general findings, concrete opinions of laypeople must be obtained for a potential implementation. This raises the questions in which situations, to what extent and with which interaction possibilities explanations must be designed in order to have a benefit for laypeople. Since the uncertainty of humans and AIs also seems to be an important factor, the aim is to clarify under which conditions users believe they are following a decision by an AI. The development of Explainable AI should aim to increase trust in the systems, so it is also important to ask about factors that promote trust in the opinion of laypeople. On the psychological level, it is interesting to see how an explanation affects people, which psychological characteristics of people are addressed and how they feel about a decision. Especially the last aspect could lead to the realization that explanations are useful even if they do not influence decisions at all. Namely, when a person, through an explanation, has a positive feeling after a decision. This study seeks to explore these issues in order to gain insights from a layperson's perspective and to provide further directions for research.

3. Methods

In this section the study design, the recruiting process and the distribution of participants are described.

3.1. Study Design

The study was conducted in the form of an online survey. The selection of questions is based on questions already posed in literature and findings obtained by others as mentioned in Related Work. Additional questions are integrated for further exploratory investigation. The full set of questions can be found on GitHub[1]. The questions can be divided into six areas.

- *Importance of explanations*: the focus of this area is on the questions of whether and under what conditions an explanation is important (e.g. "It is important to me to know how an AI arrived at a decision or recommendation").
- *Extensiveness of explanations*: here, the focus is on the level of detail of the explanations, e.g. how is the interplay between coarse and detailed explanations as well as whether global or local explanations are required (e.g. "It is important to me to be able to understand the explanation of how an AI came to a decision, at least in principle, down to the smallest detail").

[1] https://github.com/ThorstenFooBar/xai-criteria-survey

- *Interactivity*: are interactive explanations desired and how could they be designed (e.g. "For a good explanation, it is important for me to be able to try out *what if* cases").
- *Certainty of human and AI*: what is the role of uncertainty or certainty of both AI and humans interacting with explanations in decision-making (e.g. "For me, it is particularly important to recognize how certain an AI is in making a decision").
- *Trust*: what factors need to be considered to design trustworthy AI (e.g. "For me, it is important for trust in an AI that it discloses its internal workings, even if I would not understand everything").
- *Self-determination*: what human characteristics, such as the need for control, are addressed by Explainable AI (e.g. "When using an AI system that provides comprehensible explanations for decisions, I would feel more competent in using the AI system.").

Each area consists of a set of questions measured on a five-item Likert scale, ranging from 1 (I do not agree at all) to 5 (I fully agree), to systematically capture the participants' opinions on the different areas. Four textual questions were added to explore in which specific situations explainability plays a role, which interactive elements are desired, and how an explanation must be designed to promote trust or harm it.

3.2. Participants

Participants were recruited in German-speaking countries, reaching $n = 103$ individuals (60.6% male, 39.4% female). The median of the age of the participants is in the group 35-49 years. 67.3 % of the participants are between 25 and 49 years old. 101 participants are employed. Two are students. 51% work in the IT sector, 6.7% in education. The remaining participants are spread across a wide variety of industries (e.g. financial industry, research, culture, medical sector etc.). At least three-quarters of the participants regularly use apps with AI-supported functions (5.0 median and 4.37 mean value on a five-item Likert scale).

4. Results

In this section the analysis results of the $n = 103$ responses are presented. For the analysis of the questions on the five-item Likert scale, descriptive statistics (mean, quartiles, standard deviation etc.) were determined. Text responses are analyzed with manual clustering and frequency analysis to extract important characteristics of trustworthy explanations. The impact of certainty and uncertainty in combination with explanations is tested for significance with the Mann-Whitney U-test.

4.1. Important Criteria for Explainable AI

In order to extract the most important criteria for Explainable AI to laypersons, the responses were sorted by first and second quartile as follows: For strong positive statements, items were selected whose first quartile was greater than or equal to 4.0. For strong negative statements, the items were selected whose first quartile is equal to 1.0 and whose second quartile is less than or equal to 3.0. The statements have been translated from German.

4.1.1. Positive Statements

The 20 statements shown in Table 1 to Table 6 emerged as strong positive responses, grouped by the categorization described in Study Design and sorted in descending order by their mean.

As can be seen in Table 1, explanations in critical and non-critical situations are of great importance for participants, with critical situations being rated higher. More important than explanations in non-critical situations, however, is the comprehensibility of the decision-making process of an AI.

Table 1
Positive statements about importance of explanations with mean value and standard deviation on a five-item Likert scale.

Importance	μ	σ
When using AI in critical situations, it is important to me that the AI can explain how its recommendations or decisions were arrived at.	4.53	0.89
It is important for me to know how an AI has come to a decision or recommendation.	4.35	0.79
I would like to know how recommendations on the internet (Netflix, Amazon, Spotify, Google search results etc.) came about.	4.08	1.01

With regard to the extensiveness of the explanations (Table 2), roughly granular explanations are desired, which can be examined in more detail if necessary.

Table 2
Positive statements about extensiveness of explanations with mean value and standard deviation on a five-item Likert scale.

Extensiveness	μ	σ
An explanation shall be given very coarsely at the beginning and presented in more detail when needed.	4.21	0.99

On the level of interactivity (Table 3), participants would like to have the opportunity to try out what-if cases in order to understand an explanation, especially in critical situations. In addition, there is a desire for the explanation to learn adaptively from user needs and become increasingly individualized.

At the level of certainty (Table 4), it is important to the participants that an AI indicates how certain it is about the decision. Even if the AI is uncertain about the decision, it should provide an explanation.

The most important point for trust in explanations and AI (Table 5), according to participants, is the need for a decision to ultimately always be made by the human. Furthermore, the explanations should be plausible and the internal working methods transparent in order to increase trust. The aforementioned what-ifs are also considered very important for trust.

Participants also assume that explanations will have a positive effect on various areas of self-determination (Table 6). They believe that they would feel more competent in dealing with an AI, that they would be motivated to use an AI that provides explanations, and that their

Table 3
Positive statements about interactivity of explanations with mean value and standard deviation on a five-item Likert scale.

Interactivity	μ	σ
You have to make a difficult decision (you could lose a lot of money, for example). You are supported in this decision by an AI. How strongly does the following statement apply to you: "I would try out a lot of "what if" cases to come to a decision".	4.27	0.92
For a good explanation, it is important for me to be able to try out "what if" cases.	4.14	0.99
An AI system whose explanations I can adapt to my own needs through continuous interaction with the system (e.g. through feedback or natural language dialogue) would be optimal for me.	4.01	0.95

Table 4
Positive statements about certainty and explanations with mean value and standard deviation on a five-item Likert scale.

Certainty	μ	σ
For me, it is particularly important to recognise how certain an AI is in making a decision.	4.39	0.83
If an AI is uncertain about a decision, I still want to get an explanation that I can evaluate for myself.	4.15	0.92

Table 5
Positive statements about trust and explanations with mean value and standard deviation on a five-item Likert scale.

Trust	μ	σ
For me, it is important for trust in an AI that, despite a plausible explanation, I can make the decision myself.	4.78	0.58
For me, it is important for trust in an AI that it provides a plausible explanation for a decision.	4.24	0.83
For me, it is important for trust in an AI that it discloses its internal workings, even if I would not understand everything.	4.16	0.98
For trust in an AI, it is important to me to be able to ask the AI "what-if" questions.	4.07	0.98

sense of security would increase. They assume that explanations would appeal to their curiosity and for this reason they would also use explanations in non-critical situations.

4.1.2. Negative Statements

The three statements in Table 7 and Table 8 emerged as strong negative responses, sorted in ascending order by their mean. Low mean values mean that many people disagree with the statement.

Table 6
Positive statements about self-determination theory and explanations with mean value and standard deviation on a five-item Likert scale.

Self-determination	μ	σ
When using an AI system that provides comprehensible explanations for decisions, I would feel more competent in using the AI system.	4.22	0.93
An AI system that provides comprehensible explanations for decisions would motivate me to use the AI system.	4.20	0.92
An AI system that provides comprehensible explanations for decisions would give me a feeling of security.	4.10	0.90
An AI system that provides comprehensible explanations for decisions would appeal to my curiosity, so that I would also refer to explanations in non-critical situations.	4.08	0.93

At the level of importance (Table 7), there is a denial by the participants that an AI should not provide an explanation if it is uncertain. This result is congruent with the above finding that an AI should provide an explanation even if it is uncertain. It is also denied that an AI does not have to provide explanations as long as the results are good. Conversely, this means that an AI should provide explanations even if it produces good results.

Table 7
Negative statements about explanations selected by first quartile equals 1.0 and second quartile lower than or equal to 3.0 grouped by categories introduced in Study Design

Importance	μ	σ
If an AI is uncertain about a decision, I don't need an explanation because I will reject the decision in any case.	2.23	1.20
As long as the AI delivers good results, I don't care how they came about.	2.27	1.16

At the interaction level (Table 8), participants would not be willing to try out very many what-ifs in non-critical situations.

Table 8
Negative statements about explanations selected by first quartile equals 1.0 and second quartile lower than or equal to 3.0 grouped by categories introduced in Study Design

Interactivity	μ	σ
You have to make an easy decision (e.g. choosing a film for the evening's television). You are supported in this decision by an AI. How strongly does the following statement apply to you: "I would try a lot of "what if" cases to come to a decision".	1.99	0.97

4.2. Textual Feedback on Trustworthy Explanations

Participants were able to express how explanations for AI systems would have to be designed in order to be able to trust them. In addition, they were asked how explanations would have to be designed that could not be trusted. The answers were combined into categories based on the same content. The categories were defined manually based on the frequency of their mention. The results were then sorted by frequency of occurrence to get important characteristics. The most frequent characteristics for trustworthy explanations (with the number of mentions on y-axis) are shown in Figure 1.

Participants want transparent, comprehensible and short explanations so that they can trust them. The decision-making process should be transparent, as well as the data used for the decision. Logical plausible explanations, to be given in more detail if required, are desired for trust. Examples should also help, as well as the AI certainty already shown. In addition, several participants would like to see an indication of sources, although it was not specified which sources were meant. These could be sources of the data, the algorithms used, the training pipeline of the models and many more.

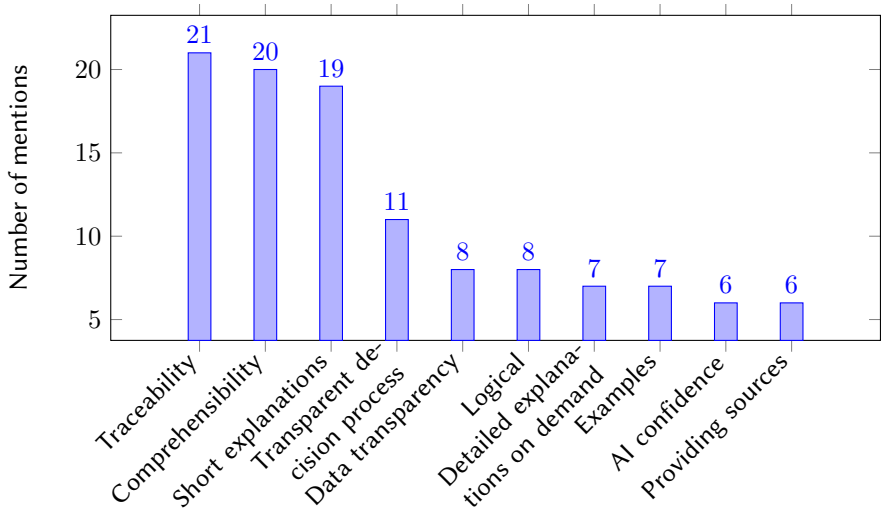

Figure 1: Most frequently mentioned characteristics of trustworthy explanations.

The most frequent characteristics of an explanation that is not trustworthy are shown in Figure 2. Not surprisingly, reversals of previously positive statements are strongly represented. These include incomprehensibility, non-traceability and the length of explanations as well as complicated, implausible and illogical implementations. Furthermore, statements are made here that attempts at manipulation through explanations would lead to a loss of trust.

In addition to these quantitative results, it can be seen on a qualitative level that explanations must be created in a human-centric way. Participants demand that explanations are truthful and honest and that assertions not simply be made. Information must not be omitted. The AI's explanations must be controllable. They should be based on scientific and verified principles.

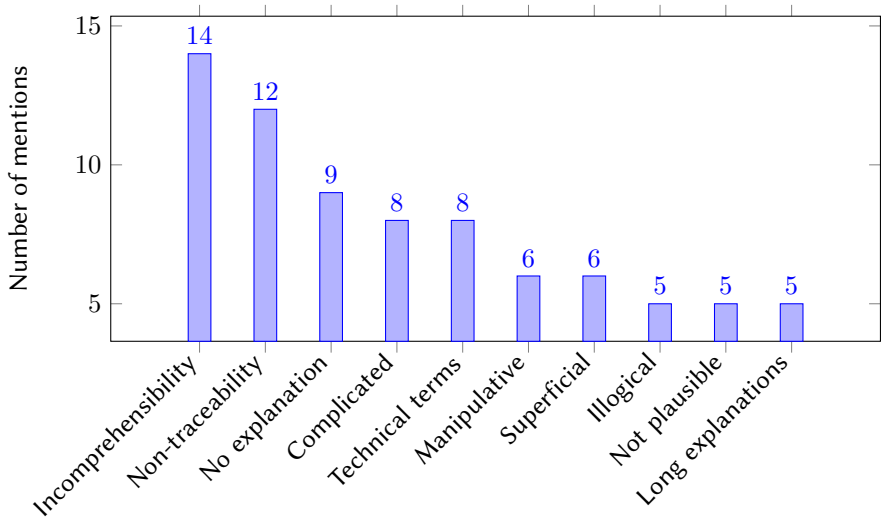

Figure 2: Most frequently mentioned characteristics of non-trustworthy explanations.

They should be neutral and a contact person should be indicated. Manipulation by the explanations must be excluded. The explanation must not serve any marketing purpose, should not promote any purchases, must not be augustly nice and must not be promoted by any sponsor. The participants would like to be able to influence the AI decision and the explanations. It should be recognizable how the user behavior influences the AI decision. User preferences must be taken into account and the deletion of data must be possible.

4.3. Textual Feedback on Interactivity

Participants were asked about the design of interactive elements for explanations. The answers were combined into categories based on the same content. The categories were defined manually based on the frequency of their mention. The results were then sorted by frequency of occurrence to get important characteristics. There were 41 answers given, as this question was not a mandatory question. The most frequent answers are shown in Figure 3.

What-if explanations followed by interactive diagrams in general are considered important interactive components for promoting trust. More specifically, explanations should be able to be compared with each other and the AI's decisions should be changeable through interaction. Context sensitivity includes statements on the adaptivity of the explanations, so that they can adapt to the user's needs. There was feedback on the interaction with a specific explanation in general, i.e. comparing, exploring and rating explanations. In addition, feedback plays a major role. The AI should be able to be improved through feedback, alternative explanations should be able to be weighted. It can also be seen that access to historical explanations and decisions is desired.

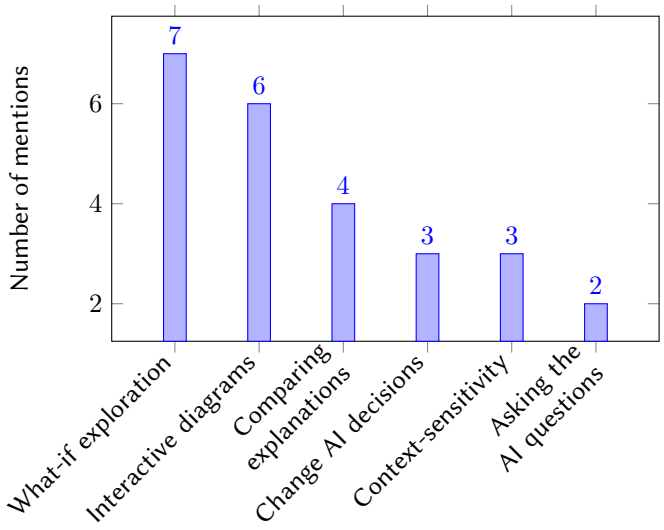

Figure 3: Most frequently mentioned interactivity characteristics.

4.4. Explanations under Uncertainty

It is reasonable to assume that uncertainty in the context of AI contributes to trust. Either in a positive or negative way. To understand the impact of uncertainty, in both human and AI, on decisions in combination with explanations, participants were confronted with several decision situations and had to indicate to whom they would leave the decision: the human, the AI or no decision at all. All the decision-making situations were in the form of *You have to make a decision that you are not very sure about. However, you have a tendency. An AI has a different tendency and is 90% sure that this decision will produce a positive result. What do you do?*. After each decision, the participants were asked how they feel about their decision on a five-item Likert scale with the range 1 (I have a very bad feeling) to 5 (I have a very good feeling). The cases where both human and AI are uncertain and have the same tendency for the decisions can be found in Table 9. Cases where the AI is certain and the human is either uncertain or certain and both parties have different tendencies for the decisions are shown in Table 10. The first three rows of the tables are the case number and the certainty levels of human and AI. In the fourth row is stated if an explanation for the AI decision is available or not. The next three rows show the distribution of the participants' decisions. The last row holds the mean values and standard deviations of the feelings about the decision made.

4.4.1. Human and AI with the same decision tendencies

In the case where the human and AI have the same tendencies, only the case where both parties are uncertain is of particular interest, as it is the worst case. As soon as one of the two parties is more confident in the decision, given the same tendency, the decision itself will only be strengthened. Table 9 shows these cases, without and with explanation.

Table 9
Percentage distribution of participant decisions in situations with uncertainty (both human and AI) with or without explanations given. The last row shows mean value and standard deviation of feelings on a five-item Likert scale.

Case	1	2
Human certainty	uncertain	uncertain
AI certainty	uncertain	uncertain
Explanation	no	yes
Human decision	75.7%	71.3%
AI decision	75.7%	71.3%
No decision	24.3%	28.7%
Feeling (μ, σ)	3.44 1.10	3.72 0.98

Table 10
Percentage distribution of participant decisions in situations with uncertainty and certainty, with or without explanations given. The last row shows mean value and standard deviation of feelings on a five-item Likert scale.

Case	3	4	5	6
Human certainty	uncertain	uncertain	certain	certain
AI certainty	certain	certain	certain	certain
Explanation	no	yes	no	yes
Human decision	15.5%	8.7%	59.2%	20.4%
AI decision	42.7%	85.4%	15.5%	54.4%
No decision	41.7%	5.8%	25.2%	25.3%
Feeling (μ, σ)	3.29 1.08	3.89 0.96	3.46 1.15	3.52 1.03

The percentage values are the same for human decision and AI decision as both have the same tendency. It can be seen that as long as human and AI are both uncertain about the decision, the human decision is preferred by roughly three-quarters of the participants, regardless of whether an explanation is given or not. Explanations do not seem to matter in these situations.

4.4.2. Human and AI with different decision tendencies

Table 10 presents cases where the AI is certain about its decision. Human certainty varies, as does the presence of an explanation. The AI and the human have different tendencies regarding the decision.

It can be seen that the decision to rely on the AI is influenced by the presence of an explanation (transitions from case 3 to 4 and from case 5 to 6). It is worth noting that 54.4% of the participants would rely on the AI even if they are confident about their own decision preference and have a different tendency compared to the AI if an explanation is present.

Table 11
U_1, U_2 and p values of the Mann-Whitney U-test. Significant results for cases (1,2) and (3,4) but not for case (5,6).

Cases	U_1	U_2	p
(1,2)	4515.5	6093.5	< 0.05
(3,4)	3558.0	7051.0	< 0.05
(5,6)	5172.0	5437.0	> 0.05

4.4.3. Feelings about the decisions

A Mann-Whitney U-test was used to examine whether participants felt significantly better about their decision with an explanation in place. The U-test was performed on the mean values between case pairs (1,2), (3,4), and (5,6) with a significance level $\alpha = 0.05$. Table 4 shows the U_1, U_2 and p values of the U-test.

It can be seen that the null hypothesis can be rejected for case pairs (1,2) and (3,4) but not for (5,6). In the case where the human and AI are both uncertain and have the same tendencies for the decision, having the AI provide an explanation very likely leads to a more positive feeling about the decision made. Although, the choice is still the same. The case where the human is uncertain, but the AI is confident in the decision, and they have different tendencies for the decision, very likely leads to a more positive feeling to choose the AI decision if the AI provides an explanation. If the human is also confident in his/her decision, choosing the AI decision does not lead to a more positive feeling. Although, it is becoming more likely that the human will choose the AI decision.

5. Discussion

The results of the study show that participants demand explanations for AI decisions. This applies to both critical and non-critical situations. Even when the AI delivers very good results, explanations are desired. This is also true when the AI is uncertain. It is an interesting question which human factors play a decisive role in the respective situations. In non-critical situations, it can be assumed that curiosity is one of the driving factors. In critical situations, it may be the need for competence as well as security. However, these assumptions still need to be confirmed empirically in further experiments. If explanations have such a high importance for AI decisions, it is necessary to explore the exact impact of explanations in critical and non-critical situations on trust.

It has been shown that explanations on a coarse-granular level are useful at the beginning and can be explored in detail only when needed. This raises the question of how much of an impact this approach has on trust. The exact relationship between coarse-granular and detailed representation with regard to trust must also be investigated empirically in experiments in the future. It could be determined that participants imagine that explanations should be examined intensively in critical situations. However, many critical situations are also time-critical (e.g., autonomous driving). In these situations, extensive analyses cannot be performed. It follows that it is necessary to explore how granular the explanations have to be chosen depending

on the specific situation. Global explanations describing the internal workings of AI may be important in initial trust building. Continuous explanation of individual decisions probably contributes to building and maintaining trust. However, these relationships also need to be confirmed empirically first. One challenge with global explanations and the related expressed need of participants for a transparent decision-making process is that even if the inner workings of the AI are shown, humans may still not be able to understand them because AI models work differently than human decision-making processes. The question here is how global explanations can be presented in an appropriately simple way and which of them have a positive impact on trust.

Interactive elements are considered an important factor of a good explanation, which may also be related to the social nature of explanations expressed by [5]. What-if exploration and interactive diagrams are considered important implementations. Participants would be willing to try many alternative cases in critical situations. In non-critical situations, a what-if exploration would not be conducted as intensively because those explorations are more driven by the curiosity of the participants. This raises the question of how much time people would actually put into exploration and under what conditions it would be interrupted. In addition, it remains open how much impact what-if explorations and interactive diagrams have on trust.

It could be shown that certainty/uncertainty of humans and AI together with explanations play a role in the choice of a decision. There are situations of certainty and uncertainty in which explanations positively influence the feeling about a decision, and there are situations in which explanations lead to a change in the decision toward reliance on AI. If humans rely on the decision of an AI, this can be seen as an expression of trust. In further experiments, the insights gained in this work need to be explored in concrete situations. The question is how much of an impact does the communication of certainty/uncertainty along with providing an explanation have on trust. In addition to uncertainty considerations, it is important that people can make the decision themselves. The AI must provide a plausible explanation and even offer the possibility to explore explanations and compare them with other, also historical explanations.

References

[1] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3145–3153.

[2] M. T. Ribeiro, S. Singh, 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, 2016, pp. 1135–1144.

[3] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: International Conference on Parallel Problem Solving from Nature, Springer, 2020, pp. 448–469.

[4] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously., J. Mach. Learn. Res. 20 (2019) 1–81.

[5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988. doi:https://doi.org/10.1016/j.artint.2018.07.007.
[6] M. Ribera, À. Lapedriza, Can we do better explanations? a proposal of user-centered explainable ai, in: IUI Workshops, 2019.
[7] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models., ITU Journal: ICT Discoveries Special Issue No.1 (2017) (2017).
[8] H. P. Grice, Logic and conversation, in: Speech acts, Brill, 1975, pp. 41–58.