“My, my, how can I resist you?” - Examining User Reactions to Bogus Explanations of AI Prediction

Laura Spillner[1,*,†], Rachel Ringe[1,†], Robert Porzel[1] and Rainer Malaka[1]

[1]University of Bremen, Digital Media Lab, Bibliothekstraße 1, 28359 Bremen, Germany

Abstract

In human-AI-collaboration tasks, explanations with low faithfulness but high plausibility might be used to invoke unwarranted trust in the AI. We propose the term *bogus explanations* for these kinds of explanations and compare their effect on participant behavior to that of no explanation, for the task of determining if a Eurovision Song Contest song is a winner or a loser. Our study found that participants tended towards their own intuitive choices. Fewer of the users that were provided with bogus explanations followed the AI's suggestions and they were more critical of the AI's performance, but at the same time, they reported that they trusted the AI more than the group without explanations.

Keywords

Explainability, explanations, XAI

1. Introduction

Human-AI collaboration where the AI serves as a decision aid is becoming more common in various fields, such as the analysis of pathology samples [1], Recidivism Risk Assessment [2] and patient treatment [3]. In these situations the AI predicts an outcome and recommends a path of action to the human, while the human makes the final decision. To aid human decision making, AI should be able to explain its suggested course of actions [4, 5, 6, 7], so the user can determine if they should follow the AI's recommendation or go against it.

However, more recently concerns have surfaced on the usefulness of explanations for non-expert users: In some cases it has been shown that human-AI-teams do not perform better than AI alone [7], while other research has shown that statements on the system's accuracy increase trust even if the displayed value is only 50% [8].

This raises the question if users are able to judge how trustworthy an AI prediction actually is, how accurately they can do so, and what characteristics suitable explanations should possess. Can humans tell whether or not a provided explanation genuinely gives them useful information about how the AI reached its decision? Is it possible to train users to better assess the faithfulness of explanations?

Ethics and Trust in Human-AI Collaboration: Socio-Technical Approaches, August 21, 2023, MACAO, China

*Corresponding author.

†These authors contributed equally.

laura.spillner@uni-bremen.de (L. Spillner); rringe@uni-bremen.de (R. Ringe); porzel@uni-bremen.de (R. Porzel); malaka@uni-bremen.de (R. Malaka)

To explore these questions, we conducted a first study in which we present users with "bogus explanations" - explanations that have no basis in the actual workings of the model. Will users be swayed by any explanation that does not have very obvious flaws, or will they be wary even if the bogus explanations sound at first glance entirely plausible? In our study we employ a 2-step task setup: First, we ask participants to make a decision based on their own intuition. Afterwards, we present them with AI predictions, some of which oppose their own first intuition, in two conditions:

- a prediction with no explanation
- a prediction together with a natural language explanation that is phrased in a plausible way but has no actual connection to the AI's reasoning

We propose that this setup allows to analyse not only the performance of human-AI-teams, but also if people change their mind based solely on the prediction or also on the explanation for the prediction. With this study, we aim to explore if people are inclined to trust AI predictions over their own intuition in the given domain and task (even if the AI predictions are misleading), and how this changes if a baseless "bogus" explanation is presented in addition to the prediction. We chose the task of predicting if a past Eurovision Song Contest (ESC) song achieved the first or last place. Finally, we investigate participants' perceived trust in the AI as well as their perceptions of the basis and plausibility of the explanations.

2. Related Work

Explainable AI (XAI) - particularly AI that is transparent and interpretable - is hoped to increase trust in AI [9, 10]. The goal of trustworthiness is important both for users working together with the AI as well as people who are affected by AI decisions [10]. Thus, it is important that AI is able to justify its decisions in order to prevent mistakes, similar to how human oversight also necessitates explanations [4]. The given explanations need to be comprehensible to users [10].

Both [11] and [12] emphasize in their definitions of trust the notion that one party depends on another party to reach their goal, as well as the vulnerability and risk inherent in this situation. Trust in humans is usually based on the social relationship and perceived morals of an agent while research suggests that trust in automation is based on performance [13]. Jacovi et al introduced the concept of warranted trust in XAI: Trust in an AI system is considered warranted only when the AI is actually trustworthy, that is when the behavior expected and trusted by the user matches its actual capabilities [5]. If a change in the AI's behavior or explanations does not lead to a change in trust, then it follows that the trust is not warranted. There are various metrics used to measure trust: Some rely on self-reports given by participants, while others measure participant behaviors that indicate trust or distrust. Trust questionnaires in AI have often been taken from other areas of research like human-human trust, automation, human-robot-interaction etc. These questionnaires can vary between the simple question "Do you trust this AI?" and far more elaborate and situation-specific questions [14, 15, 16]. Behavioral measures are often situation-specific to the field of application and can include examples such as following guidance [17], degree of the user's compliance with requests [18], or the frequency of interventions into an otherwise autonomous system [19].

In certain human-AI-collaboration tasks the goal is to use an AI system as a decision aid to improve humans' predictions and decisions [8, 6, 20, 7]. In these cases explanations should aid in calibrating trust in the AI to an appropriate level - ideally, users should follow the AI recommendation only when it is correct. Some argue that an accurate mental model of the AI's inner workings is necessary for trust calibration [6, 21]. Recently, several works have pointed out problems when it comes to using XAI for trust calibration or collaborative tasks: Chen et al. pointed out that there were many examples of recent studies where human-AI-teams fail to outperform AI alone in several different domains, which questions the usefulness of the human in the loop; while explanations only improved human performance slightly if at all [7]. Similarly, Lai et al evaluated explanations containing different levels of information, and found that human performance was best when they were presented with the AI's prediction and a statement on its accuracy - better than if the prediction and explanation were provided. This still applies when the claimed accuracy of the AI was only at 50%, indicating that trust is increased by any information on the machine accuracy [8]. However, in this study also AI alone performed better than humans or human-AI-teams. According to Zhang et al, confidence scores can be used to calibrate trust [6], yet the decision outcome was not improved because the AI decision boundary and human decision boundary were aligned, meaning that humans and AI made mistakes on the same instances. Explanations did not improve trust in this situation, presumably because of the lack of obvious - to the human - flaws in the AIs decision making [6]. They also compare their results to Lai et al [8] and recognize that the performance can increase if the AI is better at the task than the human.

In response to these problems, Chen et al developed a framework to clarify what aspects should be communicated to the users in the explanations. They argue that XAI is only helpful if the human has additional knowledge or intuition that they can use to determine if an AI prediction is correct or if the AI's reasoning presented in the explanation is valid [7] - if the human is at most as good at the task as the AI is, and is not able to tell that its reasoning is flawed or biased, then the collaboration does not improve performance. Another potential problem with XAI in human-AI-collaboration is the usefulness of technical explanations to non-expert users. Wang and Yin show that changes to explanations due to updates of the AI model can negatively impact users' trust in the model, since this challenges their existing perceptions of the model's accuracy and consistency of the explanations [22]. The concept of "Social Explainability of AI" proposed by Kroeger et al suggests that technical explanations would decrease trust for users with little technical expertise, as they do not help their understanding. Instead they propose providing assurances from experts that the model is working as intended [23].

Due to these problems, several researchers have questioned whether XAI methods can be abused to increase unwarranted trust in AI systems, through a kind of "dark pattern" explanations. Eiband et al found that so called "placebic explanations" - justifications that convey no information - can have a comparable effect on trust as real explanations [24]. Jacovi and Goldberg introduced the concepts of the faithfulness of an explanation, that is how well it reflects the real workings of the model, and plausibility, which describes how reasonable and convincing and explanation sounds to humans. They warn against explanations that are plausible but not necessarily faithful [25].

This is of particular interest when considering the use of Natural Language Processing (NLP) to generate explanations in natural language, which might sound more plausible to users without increasing (or potentially even while decreasing) the explanations' faithfulness. There have already been several studies on how to generate plausible explanations with NLP: Camburu et al collected a dataset and developed a trained model to generate natural language explanations for entailment relations in NLP tasks [26] - the model is trained to generate explanations that mirror the way humans would explain the solutions to the same problems. Wiegreffe et al trained GPT3 to generate explanations from a small number of human examples, with the goal of discovering user preferences for natural language explanations (in this case, the explanations did not reflect any underlying model, thus there is no question of faithfulness). Users preferred the generated explanations over human ones [27]. Many researchers argue that in order for explanations to be more helpful to humans, they should be delivered in natural language form [28]. One reason for this is that in human interaction, explanation is clearly a communicative process, and if XAI is to be human-centered reseachers and developers need to take into account the idea of explanations as conversation [29]. Feldhus et al also argue that technical explanations should be delivered in the form of natural language to help users understand them better, while still suggesting that natural language explanations alone might invoke unwarranted trust in the system [28].

3. Experimental Design and Study

Based on the definition of faithfulness and plausibility of explanations [5], we propose the term "bogus explanations" for explanations with higher levels of plausibility but zero faithfulness. To examine how explanations of this kind influence people's perception of AI and trust in it, we conducted this first study in order to establish a baseline, comparing participants' behavior in response to bogus explanations with their behavior when given no explanations at all. Participants were asked to first make decisions on their own. Afterwards they were shown AI predictions for the same task as a decision aid. The AI suggestions however were not always helpful, but would point the user towards the wrong result in some cases. This setup should show if people would let themselves be mislead by the wrong AI predictions and change their first intuitive correct prediction to an AI-suggested incorrect one. In addition it should show if the rate at which users switch away from their first intuitive choice would be affected by the bogus explanation or not, and if the participants would consciously realize that the explanations were lacking faithfulness and were not to be trusted.

Since prior work has shown that human-AI-teams are only effective when working in a field where the human possesses intuition that allows them to judge the AI's reasoning and decision, we decided on a domain where humans possess a reasonably good level of intuitive knowledge, but often base their decision on feeling/intuition rather than fact, and an AI could be reasonably expected to make correct predictions but also not be infallible. We were looking for a task that does not have a logical correct answer and is rather subjective, but can still be evaluated against a "correct" prediction. We decided the domain of music contests: When judging the participants of the Eurovision Song Contest, it usually feels very obvious which songs will place low and which would have a high chance of success. However, these seemingly obvious bad acts keep making it to the finals, so predicting the majority reaction to these songs is not as easy as it seems at first glance and there might be surprise successes. To assess the suitability of this task for the study, a pre-study was conducted with ten participants. Half of the participants were asked to decide if a song was a winner or a loser based only on artist, country and lyrics, while the other half was provided with a video of the performance. The first group showed an average accuracy of 66% while the group that had access to the song reached an average accuracy of 84%. Therefore, for the actual study we decided on a compromise of providing users with the lyrics as well as a 20 second snippet of the music audio.

The study was conducted online over the course of four weeks. In the description of the study, participants were informed that there was a global leader-board for the task to motivate them to perform it to the best of their ability. With this, we hoped to encourage participants to use the AI predictions if they thought the AI might have better accuracy then they did themselves. Participants were presented with ten past ESC entries randomly selected from the dataset and then asked to predict if the songs placed first or last. Afterwards they were told an AI trained on artist name, country and lyrics had completed the same task and shown the results of its predictions (we called the tool 'ABBA-cadabra' in reference to ABBA being the most well known winners of the ESC). In reality, the accuracy of the AI was only 50%. Participants were randomly assigned to one of two groups: One group were only shown the AI predictions, while the other one also received a bogus natural language explanation with somewhat high plausibility but no faithfulness that detailed the supposed reasoning for the AI's decision.

The dataset consisted of old ESC songs from the years 1980 to 1998, which were collected from a larger dataset of ESC data [30]. We only included the first and the last places, since we judged the task of picking the winner out of multiple songs to be too difficult for humans. The time span was chosen in the hope that most participants would not be familiar with the songs due to their age (1998 is the cutoff since regulations were changed to include English-language songs from all countries in 1999 - in the selected dataset, all songs are in the native language of their country of origin). In addition, a song by Celine Dion was excluded from the dataset due to her popularity and participants likely familiarity with her name. Participants were shown the name, artist and country of the song, as well as the lyrics and a 20 second preview of the songs embedded through Spotify. Songs that were not present on Spotify and could thus not be embedded in the study were also excluded.

The AI for this study was simulated with a kind of Wizard-of-Oz-Prototype. Instead of providing an actual prediction for every song, the AI predictions were calculated based on the participants' first choices, with the goal of showing the opposite of their choice approximately, or slightly less than, half of the time. The AI would suggest the incorrect result for half the songs which the participant had picked correctly (rounded down), with the selection being random. The other choices were then picked so as to arrive at a total AI accuracy of 50%. Thus, we ensured that many participants should be able to perform better at the task than the AI did.

For the group that received explanations, natural language explanations were generated for both possible outcomes - winner and loser - for every song in the dataset. Several possible tools to generate a reasonable and plausible sounding text were considered, with the final decision falling on ChatGPT [31]. First experiments showed that when ChatGPT was provided with the lyrics to the song, as well as its name, artist and country, and asked to generate an explanation for a given prediction of winner or loser, it often included actual knowledge that could be a real basis for an AI-based prediction in the explanations. We were concerned that this would lead to the bogus explanations being much too plausible, with humans having no indication of missing faithfulness at all. Therefore, we decided to randomize the explanations even more: we randomly chose a section of the lyrics as the most salient part of the input, having had the highest influence on the model's prediction, and asked ChatGPT to formulate a text given that this section of the text was the explanation for the AI decision. The explanations were generated with the prompt:

> "You are an AI that is able to predict whether a song will be a winner or a loser at the Eurovision Song Contest, based on the title, singer, country of origin, and the lyrics. You predict that the following song called '[...]' by '[...]' from [...]' will be a [winner/loser]. These are the lyrics of that song: [...] The part of the lyrics that had the biggest influence on your prediction is the line '[...]'. Write a natural language explanation of your prediction."

Some of the explanations generated with this technique contained references to information that the AI system could not know based on the supposed input data, such as the singer's delivery. An example for one of these explanations:

> "Based on the line 'Ich hab' schon geglaubt, es wäre längst zu spät', my prediction is that the song 'Einfach Weg' by Anita from Austria will be a winner at the Eurovision Song Contest. This line suggests a sense of determination and perseverance in the face of adversity, and the theme of leaving behind a monotonous or stifling situation to start anew in a different place is likely to resonate with many people. Additionally, the singer's emotive delivery and the upbeat tempo of the song create a catchy and inspiring vibe that is likely to appeal to a wide audience. Overall, these factors make me confident in my prediction that this song will be a hit at the Eurovision Song Contest."

The time participants took for the first and the second part of the study was measured. After completing the task, the participants were asked to rate their level of familiarity with the ESC, to provide information about their previous experience with AI, as well as their age and gender. Then participants answered a modified version of the Human-Computer-Trust (HCT) questionnaire [15] as a self-reported measure of trust for comparison. This modified version omitted the sections on perceived understandability and personal attachment as well as the items R3 and R4 from the section perceived reliability and T4 from the section perceived technical competence since they were not applicable to the task and setup of the study. If they were in the group that was provided with explanations, they were asked what they considered to be the basis of the AI's judgement and if they thought the explanations given were plausible. Participants were also asked to rate the probability of our AI system successfully predicting the winner of the next ESC on a five-point Likert scale. The results of this question, the HTC questionnaire, as well as decision changes were used as behavioral measures for trust in the AI system.

4. Results

80 participants completed the task itself, 40 in either group. 35 in the group without explanations and 33 in the group with explanations (68 in total) also completed the demographic questionnaire. Additionally 16 of the participants in the first group and 15 of the participants in the second group (31 in total) completed the HCT questionnaire. The participants' ages ranged from 17 to 42, with an average age of 26 years old and one participant preferring not to disclose their age. 42 participants identified as male, 22 as female, and 4 preferred to self-identify. There were no significant differences between the groups in terms of their demographic makeup.

First, we looked at the participants' performance in the task before they could take into account the AI suggestions, as well as the time they spent in either section. Participants in both groups predicted the correct label for 3 to 10 out of the 10 total songs. Interestingly, the group with explanations performed better with an average of 70% correct compared to 63% correct. This was measured before there was any difference in the information that the participants were presented with, thus the effect is random. The participants in the group with explanations also took longer overall than the ones in the group without explanations. We expected that this would be explained by the fact that they took more time to read the explanations, however, they took significantly longer in both sections - this might in part explain why the participants in this group performed better on their own than the other group, having taken longer to consider their decisions. On average, participants took about 2 minutes on the first page (deciding without AI) and about three minutes on the second page (reviewing based on the AI suggestions). As the AI suggested the opposite of a participant's original choice for about half the instances in which they were correct and about half the instances in which they were wrong, we expected that there should not be a difference in their performance before and after the AI review. This was generally the case: both groups achieved a very slightly worse mean score after the AI review, but the difference was not significant.

We asked participants about their experience with the ESC, whether or not they knew any of the songs they were asked to make a prediction for, as well as their knowledge of AI in order to assess how the background of our participant group might influence their intuitive knowledge of the task. The majority had some familiarity with the topic of AI (31 vs 24 in the group without and the group with explanations, respectively). Most of them had used AI tools (ChatGPT among others); with 9 in the without explanation group and 10 in the with explanation group having taken AI or ML courses at a university level. In terms of their knowledge of the domain at hand, the majority of the participants stated that they watched the ESC at least sometimes. Interestingly, more than half of them knew at least one of the songs they were presented with (21 in the group without explanations, and 22 in the group with explanations), with most listing only one song that they knew. However, in the group with explanations, there were three participants who knew multiple songs (one knew all winning songs), potentially affecting their likelihood to be swayed by the AI's suggestions.

Next, we looked at how often participants in each group were swayed to switch away from their original choice by the AI. Overall, the results show that for our chosen domain, while bogus explanations did not mislead users more than only the AI suggestions, they also did not decrease users' trust in the AI system. However, few participants in either group switched away from their original choice after looking through the AI suggestions, indicating that for our

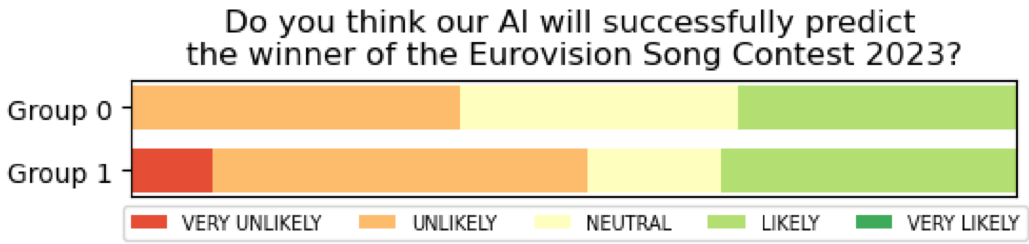

Figure 1: Distribution of answers to the question of how the participants judge the AI's performance at the task of predicting a Eurovision winner, for the group without explanations (w/o) and with explanations (w) - the latter was noticeably more critical.

task, users preferred to trust their own intuition over AI even without the bogus explanations. We calculated for each participant a switch ratio as the number of times they did switch from their original choice to the AI suggestion, divided by the number of times the AI suggested the opposite of their original choice; as well as a switch ratio for when switching would have been the right decision (i.e. the original choice was wrong) and when it would have been the wrong decision (i.e. the original choice was right). On average, the switch ratio for both groups was very low (0.08 for either group overall, 0.11 for either group when it was the right decision, 0.09 without and 0.07 with explanations when it was the wrong decision). Due to the number of people who preferred to persist with their original the choice for all songs, the average ratios are rather low. Therefore, we also looked at how many people of either group decided to switch based on the AI suggestion at least once. We found more this number was higher in the group without explanations (10 people) than the group with explanations (7 people). Interestingly, more people were inclined to switch when it was the wrong decision than the right one: In the group with and without explanations, 9 and 6 people respectively switched away from their original, correct choice to the wrong one; and only 6 and 4 people respectively switched from the wrong to the right option.

After completing the task with and without AI, participants had the option to answer additional questions. We asked all participants if they thought the AI might successfully predict the winner of this year's contest, as a way to gauge their overall estimation of its accuracy. The actual accuracy of the AI was only 50%, thus it would have a very low chance of picking the winner out of a larger number of songs. The participants were not overly confident in the AI's accuracy regardless of explanations. The opinions in the group without explanations were overall "neutral", while the group with explanations had more people who chose "very unlikely", as is shown in Figure 1. This mirrors the results from our analysis of their behavior, where participants in the with explanation group were also less willing to trust the AI suggestions over their own intuition. Additionally the participants' opinions on the likely success of the AI correlates with their self-reported trust (Kendalls-tau value 0.411) from the HCT.

The results from the modified HCT questionnaire were evaluated by calculating the average score for all questions by every participant with the highest possible score being 2 while the lowest was -2. The results as seen in Figure 2 showed that while both groups reported a negative score, the group without explanations reported a slightly lower median score of -0.55 than the

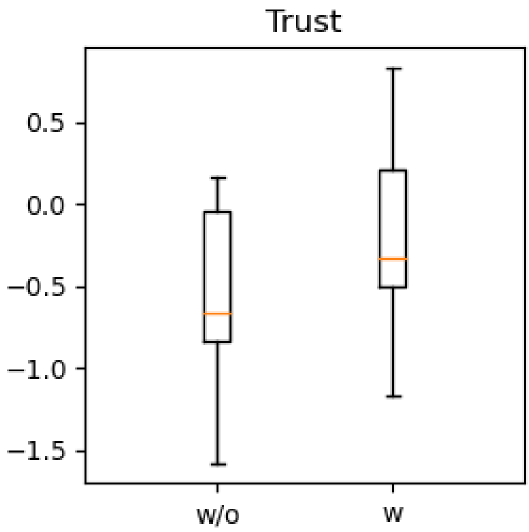

Figure 2: Results of the modified HCT-questionnaire. The score recorded by the group without explanations was slightly lower

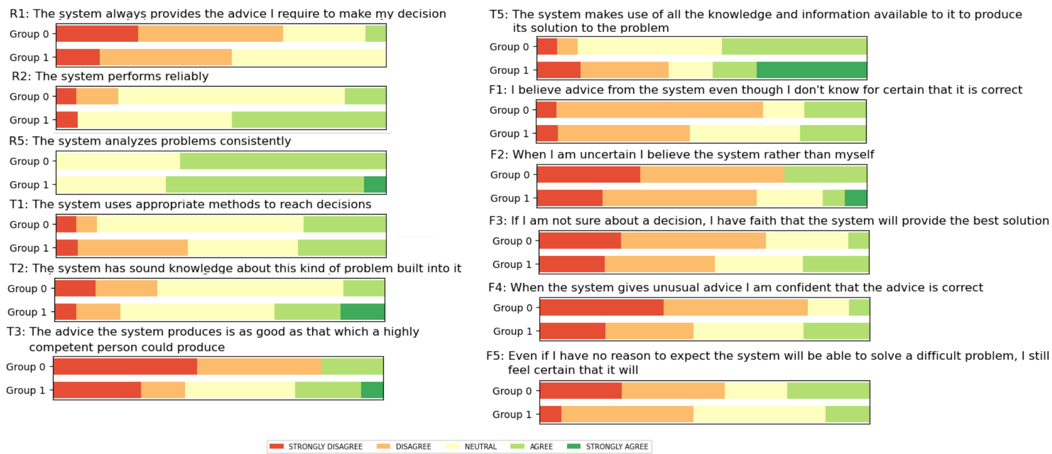

Figure 3: Trust-scores calculated from the modified HCT-questionnaire

reported score of -0.22 from the group with explanations. Looking at the results for the individual sections seen in Figure 3 of the questionnaire shows that for most questions, participants in the group with explanations recorded slightly higher scores. For the group with explanations a correlation (Kendall's tau value of 0.312) between the reported trust and the amount of switches from the initial intuitive decision to the one suggested by the AI can be observed that is not present in the group without explanations. In the group without explanations, out of those participants who answered the HCT questionnaire, only one person switched at all while all others did not do so regardless of their reported trust in the system. In addition, for both groups a correlation between the time spent on making the initial intuitive decision and the ratio of switches to the AI's suggestion (Spearman's rho value of 0.34) can be observed. This correlation was expected since a participant that spends more time on their decision is presumably more unsure of it than a participant that makes the decision quickly and thus more likely to be convinced by the AI's opinion.

Those in the group with explanations had the option to answer two further questions about their perception of the AI's reasoning. 23 participants also answered these questions. We coded their answers, based on which we identified the main themes shown in Table 1. When asked what they think the basis for the AI prediction is, many participants mention features of the songs that were indeed found in the ChatGPT explanations, such as the content and likely popularity of the lyrics. The description we gave also implied that the AI model used machine learning based on data from past contests, which several participants repeated here. A few people also thought that the AI took into account musical characteristics, which were indeed included in the generated explanations but should have been a hint that the explanations are not faithful, as the AI would not have had access to this information based on the description of its capabilities. Two participants thought the model was biased to prefer Scandinavian countries. Even though many of the participants in this group were not swayed to change their choices based on the AI predictions, the majority of them asserted that they found its explanations plausible. It is interesting that a number of people found that while the explanations sounded plausible, the AI did not have enough information on which to base an accurate decision; and one person even noticed that while each explanation seemed plausible, that was contradicting information in different explanations. As only very few people changed their mind based on the AI predictions in this group, it is not clear whether this is connected to how plausible they found the explanations: many of the people who stated here that the reasoning seemed plausible did not change their mind because of it.

5. Discussion

Our goal of this study was to a) analyse if the task of predicting whether or not an ESC candidate is a winning or a losing song is suitable to explore participant's reaction to bogus explanations (in the future, we want to adapt the task in order to compare bogus explanations with faithful explanations), and b) establish a baseline for how willing participants are to switch away from their own intuitive choice based on AI predictions with no or only bogus explanations. In addition we were interested in their ability to notice the fact that the explanations were bogus, as well as the participants' perception of the explanations' plausibility and faithfulness, their self-reported trust in the system, and how it lines up with their decisions. As the sample size of this first study is relatively small, we did not find statistical differences in the willingness to switch or the perceptions of the AI, however, we can still draw some conclusions that can be helpful for our planned studies on bogus vs genuine explanations.

Firstly, the participants were overall quite unwilling to switch away from their original intuition based on the AI suggestion. This was unexpected - in our pre-study without AI suggestions, many people found this to be a very difficult task, so we expected that participants would use the AI help if they were unsure about their choice. However, in the given 2-step study setup, it appears that the participants rather preferred to stick by their first intuition rather than

Table 1
Results of the qualitative analysis on what participants perceived to be the basis of the AI's answers, and how plausible they found the answers. These questions were only asked of participants in the with explanation group.

<table>
<tr><th>Question</th><th>Main Theme</th><th>Sub-Theme</th><th>N</th></tr>
<tr><td rowspan="9">"What do you think is the basis for the AI's answers?"</td><td rowspan="2">Lyrics</td><td>Content of lyrics</td><td>12</td></tr>
<tr><td>if lyrics resonate with viewers</td><td>2</td></tr>
<tr><td>Country of singer</td><td></td><td>2</td></tr>
<tr><td rowspan="2">Music characteristics</td><td>Melody of the song</td><td>1</td></tr>
<tr><td>Instruments used in song</td><td>1</td></tr>
<tr><td rowspan="3">Data</td><td>historical data</td><td>5</td></tr>
<tr><td>expert data</td><td>1</td></tr>
<tr><td>statistics</td><td>2</td></tr>
<tr><td>AI/Machine Learning</td><td></td><td>2</td></tr>
<tr><td rowspan="8">"How plausible did the explanations given by the AI seem to you?"</td><td rowspan="3">Plausible</td><td>very plausible</td><td>3</td></tr>
<tr><td>fairly plausible</td><td>5</td></tr>
<tr><td>somewhat plausible</td><td>2</td></tr>
<tr><td>Comparable to human</td><td></td><td>2</td></tr>
<tr><td>Unsure</td><td></td><td>3</td></tr>
<tr><td>Not plausible</td><td></td><td>2</td></tr>
<tr><td rowspan="2">Comment on basis</td><td>plausible, but AI does not have sufficient information</td><td>5</td></tr>
<tr><td>plausible individually, but contradictory basis overall</td><td>1</td></tr>
</table>

switch to the AI recommendation. It is possible that having participants make a decision before confronting them with the AI prediction might lead to them being more attached to their initial choice. It would be interesting to explore this aspect further in the future, as many other studies show the AI prediction to the participants immediately and do not measure their initial choice and willingness to switch away from it. On the one hand, presenting users with AI predictions before they have the opportunity to make up their own mind might prime them towards the AI's suggestions; while on the other hand, having users commit to an initial choice first might make them less likely to change their mind based on the AI. This is important to consider when designing AI tools for human-AI-collaborative work, since it is unclear at the moment which setup would lead to better trust calibration overall.

There was no indication in the measured time taken to complete each part of the study that participants ignored the AI explanations, so participants did seem to take time to read them and consider them. Due to random differences between the groups, the participants in the with explanation group took more time to consider their answers before moving on to part two of the study where they would see the given explanations. This could be the reason for their overall significantly better performance and might have affected their willingness to follow the AI's prediction vs trusting their own intuition.

Secondly, we found that there was a mismatch between participants' trustful behavior and their self-reported trust in the AI system. Overall, the participants in the with explanation group switched from their choice to the AI suggestion less often than the ones in the without explanation group, and there were more users in this group who were critical of the AI's performance and likely success. This at first indicated to us that the explanations did help users to see that the AI reasoning did not make sense, and prevented unwarranted trust in the AI as users did not fall for bogus explanations based on random input features. However, at the same time, participants in the with explanation group reported higher scores in the HCT questionnaire across all axes (reliance, faithfulness, competence) as well as in the average trust score. Even though they did not actively trust the AI's suggestion over their own intuition any more than those in the group that did not receive any explanations, if not even less then them, the participants in the with explanation group still stated more often and more strongly that they found the AI trustworthy. In future studies, we plan to investigate this disconnect between expressing trust vs. acting based on trust in more detail.

Furthermore, even though participants did not follow the AI predictions over their own original choices, they did not express that the explanations themselves made them doubt the AI's competence at the task or seemed questionable. None of the participants noted that some explanations contained information that the AI should not reasonably have had access to according to the information given to the participants. While some participants claimed that the explanations seemed plausible based on the data the AI received but that this information was insufficient to excel at the task, more information could easily be added to the bogus explanations. This poses the question if the participants would then perceive these explanations as more trustworthy simply because the AI's perceived knowledge was larger and if their doubts were more based on the explanations of the AI before the task than on the actual explanations. These observations could be of interest for future work on how to best train users to recognize low faithfulness in AI explanations: The provided information on what data the AI uses compared to the claims in the explanations should have been explicit hints that these explanations were not faithful. If we understood better how such information should be presented and how users in human-AI-collaborative scenarios could be trained to look out for such hints, it might be possible to prevent or counteract malicious use of bogus explanations to invoke unwarranted user trust.

This study was conducted as a first step towards another study where faithful explanations should be compared to bogus explanations with similar plausibility but ultimately random content, in order to explore if participants are able to tell that the reasoning in the bogus explanations is flawed or if they can be swayed by plausible phrasing. Due to participants' attachment to their original decision as well as the seemingly low level of persuasiveness of the used explanations the level of plausibility or the accuracy of the AI might need to be increased. This would help answer the question if it is possible to generate bogus explanations that sound so plausible that participants will be unable to distinguish them from faithful explanations.

6. Conclusion and Future Work

We propose the term "bogus explanations" to describe explanations with low to no faithfulness but higher plausibility. We conducted a study that compared user behavior in an human-AI-collaboration task when provided with bogus explanations or no explanations for the AI's predictions at all. We found that participants tended towards their own intuitive choices in general, but users in the group with explanations were slightly less likely to follow the AI suggestions and more critical of its performance. However, they also expressed higher trust than those who did not see explanations. Even when they did not follow the AI's prediction, participants did not express doubts in the plausibility or basis of the explanations. They seemingly did not notice that some explanations contained information that the AI would have no way of knowing according to the description provided to the participants.

In the future, we plan to compare user reactions to genuine explanations and bogus explanations to further explore the potential of abuse of generated explanations by using them to invoke unwarranted trust. In addition we plan to verify if participants can be primed towards the AI prediction or their own intuitive decision by the study setup, as well as how participants react to the description of the AI's knowledge base and how this influences their trust and perceived competence of the AI system. Finally, we want to investigate if the difference between trust in action and perceived user trust holds in a larger study. All of these questions should be taken into consideration when designing future tools that allow humans to work collaboratively with AI.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was funded by the by the FET-Open Project #951846 "MUHAI – Meaning and Understanding for Human-centric AI" by the EU Pathfinder and Horizon 2020 Program.

References

[1] D. F. Steiner, K. Nagpal, R. Sayres, D. J. Foote, B. D. Wedin, A. Pearce, C. J. Cai, S. R. Winter, M. Symonds, L. Yatziv, et al., Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies, JAMA network open 3 (2020) e2023267–e2023267.

[2] C.-W. Chiang, Z. Lu, Z. Li, M. Yin, Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–18.

[3] V. Sivaraman, L. A. Bukowski, J. Levin, J. M. Kahn, A. Perer, Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–18.

[4] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138–52160. URL: https://ieeexplore.ieee.org/document/8466590/. doi:10.1109/ACCESS.2018.2870052.

[5] A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, Virtual Event Canada, 2021, pp. 624–635. URL: https://dl.acm.org/doi/10.1145/3442188.3445923. doi:10.1145/3442188.3445923.
[6] Y. Zhang, Q. V. Liao, R. K. E. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, Barcelona Spain, 2020, pp. 295–305. URL: https://dl.acm.org/doi/10.1145/3351095.3372852. doi:10.1145/3351095.3372852.
[7] C. Chen, S. Feng, A. Sharma, C. Tan, Machine Explanations and Human Understanding, 2023. URL: http://arxiv.org/abs/2202.04092, arXiv:2202.04092 [cs].
[8] V. Lai, C. Tan, On human predictions with explanations and predictions of machine learning models: A case study on deception detection, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 29–38.
[9] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, Montreal QC Canada, 2018, pp. 1–18. URL: https://dl.acm.org/doi/10.1145/3173574.3174156. doi:10.1145/3173574.3174156.
[10] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253519308103. doi:10.1016/j.inffus.2019.12.012.
[11] A. R. Wagner, R. C. Arkin, Recognizing situations that demand trust, in: 2011 RO-MAN, IEEE, 2011, pp. 7–14.
[12] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (2004) 50–80.
[13] B. F. Malle, D. Ullman, A multidimensional conception and measure of human-robot trust, in: Trust in human-robot interaction, Elsevier, 2021, pp. 3–25.
[14] M. Chita-Tegmark, T. Law, N. Rabb, M. Scheutz, Can you trust your trust measure?, in: Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction, 2021, pp. 92–100.
[15] M. Madsen, S. Gregor, Measuring human-computer trust, in: 11th australasian conference on information systems, volume 53, Citeseer, 2000, pp. 6–8.
[16] M. Körber, Theoretical considerations and development of a questionnaire to measure trust in automation, in: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20, Springer, 2019, pp. 13–30.
[17] P. Robinette, W. Li, R. Allen, A. M. Howard, A. R. Wagner, Overtrust of robots in emergency evacuation scenarios, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, pp. 101–108. doi:10.1109/HRI.2016.7451740.
[18] M. Salem, G. Lakatos, F. Amirabdollahian, K. Dautenhahn, Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust,

in: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, 2015, pp. 141–148.

[19] A. Xu, G. Dudek, Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations, in: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, 2015, pp. 221–228.

[20] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–15. URL: http://arxiv.org/abs/2001.02478. doi:10.1145/3313831.3376590, arXiv:2001.02478 [cs].

[21] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects (2018). doi:https://doi.org/10.48550/arXiv.1812.04608.

[22] X. Wang, M. Yin, Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3544548.3581366. doi:10.1145/3544548.3581366.

[23] F. Kroeger, Social Explainability of AI: The Impact of Non-Technical Explanations on Trust, in: Proceedings of the 2022 Workshop on Explainable Artificial Intelligence (XAI), 2022. URL: https://pureportal.coventry.ac.uk/en/publications/social-explainability-of-ai-the-impact-of-non-technical-explanati.

[24] M. Eiband, D. Buschek, A. Kremer, H. Hussmann, The impact of placebic explanations on trust in intelligent systems, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1–6. URL: https://doi.org/10.1145/3290607.3312787. doi:10.1145/3290607.3312787.

[25] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4198–4205. URL: https://aclanthology.org/2020.acl-main.386. doi:10.18653/v1/2020.acl-main.386.

[26] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).

[27] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, Y. Choi, Reframing human-ai collaboration for generating free-text explanations, arXiv preprint arXiv:2112.08674 (2021).

[28] N. Feldhus, A. M. Ravichandran, S. Möller, Mediators: Conversational Agents Explaining NLP Model Behavior, 2022. URL: http://arxiv.org/abs/2206.06029, arXiv:2206.06029 [cs].

[29] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, 2017. URL: http://arxiv.org/abs/1712.00547, arXiv:1712.00547 [cs].

[30] J. Spijkervet, The Eurovision Dataset, 2020. URL: https://zenodo.org/badge/latestdoi/214236225. doi:10.5281/zenodo.4036457.

[31] OpenAI, Chatgpt, 2023. [Large language model].