Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems

Cecilia Panigutti*
cecilia.panigutti@sns.it
Università di Pisa
Pisa, Italy
Scuola Normale Superiore
Pisa, Italy

Andrea Beretta*
andrea.beretta@isti.cnr.it
CNR
Pisa, Italy

Dino Pedreschi
dino.pedreschi@unipi.it
Università di Pisa
Pisa, Italy

Fosca Giannotti
fosca.giannotti@sns.it
CNR
Pisa, Italy
Scuola Normale Superiore
Pisa, Italy

ABSTRACT

The field of eXplainable Artificial Intelligence (XAI) focuses on providing explanations for AI systems' decisions. XAI applications to AI-based Clinical Decision Support Systems (DSS) should increase trust in the DSS by allowing clinicians to investigate the reasons behind its suggestions. In this paper, we present the results of a user study on the impact of advice from a clinical DSS on healthcare providers' judgment in two different cases: the case where the clinical DSS explains its suggestion and the case it does not. We examined the weight of advice, the behavioral intention to use the system, and the perceptions with quantitative and qualitative measures. Our results indicate a more significant impact of advice when an explanation for the DSS decision is provided. Additionally, through the open-ended questions, we provide some insights on how to improve the explanations in the diagnosis forecasts for healthcare assistants, nurses, and doctors.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Computing methodologies** → *Artificial intelligence*; *Cognitive science*.

KEYWORDS

XAI; eXplainable AI; HCI; User Study; Behavioral intention; Trust; Advice-taking; Clinical Decision Support System

ACM Reference Format:
Cecilia Panigutti, Andrea Beretta, Dino Pedreschi, and Fosca Giannotti. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3491102.3502104

*Both authors contributed equally to this research.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3502104

1 INTRODUCTION

While some claims have been made about a future where AI will replace doctors [43], AI is more likely to become an essential tool in doctors' service, allowing them to outsource mundane tasks to algorithms and focus on more serious matters [78]. Both AI and human doctors will have complementary roles reflecting their strengths and weaknesses. Therefore, it is of pivotal importance to develop an AI technology able to work synergistically with doctors.

Current AI technologies have many shortcomings that hinder their adoption in the real world. One of the most prominent examples is the *black-box* nature of most state-of-the-art AI systems. Indeed, these models might have millions of parameters, capturing the extreme nonlinearities of the input features, making their internal decision-making process hard to interpret by human beings. The noninterpretability of such models makes it difficult to examine their reliability, identify potential malfunctions and prevent them from happening again. In the healthcare context, AI-based Clinical Decision Support Systems (DSS) having a black-box model at their core prevents the clinician from investigating unexpected findings and performing a differential diagnosis process.

In recent years, developing methods to explain AI models reasoning has become the focus of many of the scientific community's efforts, particularly those of the field of eXplainable AI (XAI) [36]. Many XAI methods have been developed for the most varied type of data and algorithm [35, 60]. However, only a few of these methods are tested on real users, and even fewer were designed with the end-user in mind [56]. AI explanations are of pivotal importance for adopting the model since they allow humans to build a shared mental model with the AI and increase trust in its suggestions. The study of the qualities that make a good explanation, i.e., a good interface between humans and AI, and the impact of the explanation on the user behavioral intention needs to be performed involving the end-users, i.e., the healthcare providers, and observing the use of explanation in the appropriate decisional context [7, 53].

In this paper, we present the results of an online user study on the impact of AI explanations in the medical field. For our experiment, we considered a state-of-the-art XAI method tailored to deal with healthcare data [66], and we employed a recurrent neural network to act as clinical DSS [18]. The final purpose of our research is twofold. Firstly, we aim to understand how explanations could enhance the trust in the AI system and the intention of using an AI system in the medical field. Secondly, we aim at giving suggestions and guidelines to the designers and researchers of XAI methods to increase their adoption in the medical field. Our research questions are the following:

- **RQ1**: How do AI explanations impact users' trust in algorithmic recommendations in healthcare?
- **RQ2**: How do AI explanations impact users' behavioral intention of using the system in the healthcare context?

In particular, we want to test the following main hypotheses:

- **Hp1**: Participants trust more the algorithmic suggestion when presented with the explanation.
- **Hp2**: Participants feel more confident when they use the system that provides an explanation.
- **Hp3**: Participants have a higher behavioral intention to use the system that provides an explanation.
- **Hp4**: Participants express higher trust in the system that provides an explanation.

2 RELATED WORK

EXplainable Artificial Intelligence (XAI) is a term coined in 2017 by DARPA for its homonymous program [36]. However, the study of techniques whose goal is to explain (i.e., capability to present in human-understandable terms [25]) the decision-making process of an AI system is as old as the AI field itself [57]. This topic has recently witnessed an increased interest that generated vast literature on AI transparency and explainability [8, 35]. Indeed, the popularity of such techniques matches the increasing use of *black-box* AI systems, i.e., systems whose internal decision-making process is obscure. Being able to explain clinical decisions to patients and be held accountable for adverse outcomes of their diagnosis are key ethical responsibilities of every doctor [58, 64]. Furthermore, in the EU, explicability is a legal requirement for high-risk AI applications such as the ones pertaining to health [1, 20, 30, 37, 52].

While several XAI methods have been developed in the past years, only a few considered the specific application domain. Consider, for example, two of the most popular XAI methods: LIME [68] and SHAP [51]. Similar to the XAI method employed in our experiment, they provide local explanations that summarize each feature's influence on the model outcome [21]. These two methods are *model-agnostic* and *application-agnostic*, meaning that they are able to extract an explanation from any type of black-box AI model [57] regardless of the application domain. While the *model-agnostic* approach to XAI offers great flexibility to the use of these methods, the *application-agnostic* approach implies that the specific user needs are not considered [4]. An interesting line of research is that of XAI methods that are not completely agnostic and tailor the explanations to the medical field, either by incorporating medical knowledge in the explanation process [5, 19, 66, 85] or focusing on specific healthcare data characteristics and use cases [54, 63, 65]. In our experiment, we focus on the healthcare application domain, and therefore we test one of these XAI methods. In particular, one that incorporates medical knowledge in the explanation process and summarizes the features taking into consideration their medical meaning [66].

However, even medical application-aware XAI methods rarely design the explanation with the end-user in mind. Furthermore, only a few of them tested the efficacy of their explanations on a group of health care professionals. A recent survey has shown that explanations of black-box AI models are mainly used by machine learning engineers to debug their model in the development phase [6]. Nevertheless, debugging the model is only one of the needs expressed in another recent study that analyzed the demands of transparency of several stakeholders [10]. Among those needs, building trust is particularly relevant to this paper. Trust plays a central role in the adoption of new technologies, and explanations of AI recommendations are often touted as the solution to trust issues [31, 68, 81, 83]. Ideally, explaining clinical DSS recommendations should help clinicians with *trust calibration*, i.e., properly adjusting their level of trust according to the actual reliability of the AI system [69]. There are several levels of trust falling along a spectrum ranging from complete distrust to overreliance on AI. Both extremes have been observed towards AI-based clinical DSSs. On the one hand, some works have shown that clinicians tend to over-rely on automated suggestions by taking less initiative [47] or accepting incorrect diagnoses suggested by AI [38]. This phenomenon is known as *automation bias* [46, 72] and can be particularly dangerous in critical domains such as medicine. On the other hand, physicians are reluctant to trust algorithms that they do not understand [16, 71] and might be subject to *algorithm aversion* [23], which is the human tendency to discount algorithmic advice [50]. Distrust in AI applications in medicine also comes from doctors' fear of legal repercussions if something goes wrong due to unclear liability regimes [61, 76].

While, at first glance, explanations of such DSS seem the solution to these issues, some studies suggested that explanations can be inadequate to deal with overreliance on flawed algorithms [42]. Furthermore, explanations might even increase overreliance on AI-based clinical DSS [13, 29, 44], and it might be necessary to design the system to force the user to engage in analytical thinking when explanations require substantial cognitive effort to be evaluated [12]. These findings highlight the importance of involving the end-user of the explanation when evaluating its efficacy and, ideally, in the design phase. The fact that the developers of XAI methods design explanations for themselves creates a gap between state-of-the-art XAI explanations and end-users. A few works have tried to close such a gap in the medical field by involving the doctors in the design procedure [45, 70, 83] or by performing exploratory surveys [16, 48, 77]. Despite these recent efforts, most of the research has been focused on laypeople [3, 17, 59]. However, several works have shown that users' domain expertise is relevant to the trust calibration process [32, 62, 82, 86], e.g., novice users tend to overrely on AI suggestions. For these reasons, in our study, we focus on the impact of explanation on advice-taking involving a specific pool of end-users, i.e., healthcare providers, and observing the use of explanation in the appropriate decisional context [7, 11, 27, 53], i.e., while performing a task supported by a clinical DSS.

Trust can be measured both by employing *explicit* and *implicit* measures. Explicit measures involve using trust scales that directly ask users whether they trust the AI or not [41], while implicit measures rely on operationalizing the definition of trust in terms of user behavior: does the user changes his or her behavior after receiving the AI-system suggestion? [84] Indeed, in the context of decision-making, trust is positively associated with advice-taking [33, 74]. Advice-taking can be measured using the *Weight Of Advice* (WOA) [39], i.e., the extent to which participants change their initial estimate after receiving the AI system's suggestion. Finally, another important factor to consider is the perceived *explanation quality*. Indeed, *good* explanations enable end-users to develop an appropriate mental model of how the AI system works, facilitating the trust calibration process. To measure explanations quality, we employed the *explanation satisfaction scale* [41] which measures explanations' understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness from users' point of view.

3 METHODS

3.1 Participants

We ran an online experiment on the *Prolific* platform (www.prolific.co). We prescreened participants to be healthcare providers (doctors, nurses, paramedics, and emergency services providers), fluent in English, and high approval rate, i.e. a high number of past approved submissions on the Prolific platform. All participants provided written informed consent, and the local Research Ethics Committee approved the study. Each participant was asked to perform a task (detailed below) and answer a set of questionnaires and received a compensation of 6.20£ for it.

3.2 Estimation task

To evaluate whether the explanation of the algorithmic recommendation influenced participants' behavioral intention and trust in the clinical DSS, we used an *estimation task*. During the estimation task, the participant is asked to make an estimate before and after being presented with the algorithmic recommendation. In this case, the task was to estimate the chances of a patient suffering from an acute myocardial infarction (acute MI) in the near future. Participants were first presented with the patient's clinical history and asked to make an initial estimate based on their knowledge and experience. Then they were shown the algorithmic suggestion, and they were asked to make a second and final estimate. Participants were also asked to indicate their *confidence level* after each estimate. This task allowed participants to decide how much they wanted to rely on the algorithmic suggestion, weighing it compared to their first estimate. Our paradigm adapts to the Judge-Advisor System (JAS) [73, 74]. In a JAS, there are two distinct roles in the decision-making process: the judge and the advisor. While the advisor provides suggestions and advice to the judge, the judge is the only one responsible for the final decision. This framework perfectly fits our case: the clinical DSS is the advisor, and the healthcare provider is the judge, solely responsible for providing appropriate care for the patient. Such a framework is widely used in algorithm reliance and aversion studies [24, 50].

3.3 Experimental design and collected data

The experimental design followed a two-cell (only AI suggestion vs. AI suggestion and explanation) within-subjects design. Each participant was asked to perform the estimation task twice: once using the interface providing only the AI suggestion and once using the interface providing the suggestion and the explanation. To prevent the learning effect, each participant used the two interfaces on two different yet analogous patients. To prevent order effect, participants were randomly assigned to different experimental groups to control the order of presentation of the different types of algorithmic suggestions (with or without explanation). We also controlled for confounding factors such as participants' familiarity and involvement in the task [26], demographic information such as gender, age, and the type of medical profession. We also controlled for participants' Need For Cognition (NFC) - an aspect related to the individual tendency to enjoy effortful cognitive (5-point Likert scale, from 1="strongly disagree to 5="strongly agree") [14, 55].

Our main dependent variable was the Weight of Advice (WOA). The WOA measures the degree to which the algorithmic suggestion (with or without explanation) influences the participant's estimate. Indeed the WOA quantifies advice-taking, i.e., how much the participants changed their initial assessment after the algorithmic suggestion. Advice-taking is defined as the ratio of two differences: first, the judge's post-advice and pre-advice estimates; second, the difference between the advisor's suggestion and the judge's pre-advice assessment [39]; $WOA = \frac{|F-I|}{|A-I|}$, where F and I are respectively the final and initial participant's estimates, while A is the algorithmic suggestion. The clinical DSS employed in the experiment performed binary prediction on whether a patient would have an acute MI in the near future or not, i.e., $A = 0$ or $A = 100$. Since we did not want to add the algorithm's accuracy as an additional degree of freedom of the experiment, we selected only patients correctly predicted by the algorithm as having an acute MI, therefore $A = 100$ in all cases. Furthermore, selecting only patients in need of urgent care allows creating a scenario that entails risk and, therefore, can yield insight into trust in the AI recommendation [2]. To better understand the influence of algorithmic suggestions on participants' estimates, they were asked to estimate the patient's chances of developing an acute MI on a scale from 0 to 100% rather than in the binary format of the algorithm. Participants were also asked to express their confidence in the estimate on a sliding scale. While the WOA can be considered an *implicit* measure of trust (because trust is positively correlated with advice-taking [33, 74]), we decided to measure also the *explicit* trust in the system by directly asking participants' perception of the system reliability, predictability, and efficiency (5-point Likert scale, from 1="strongly disagree to 5="strongly agree") [2, 15, 41]. We measured the Behavioral Intention (BI), or the participants' intention to actually use the presented systems. We followed the Technology Acceptance Model (TAM) [79] and the Unified Theory of Acceptance and Use of Technology Model (UTAUT) [80]. According to our purpose, we adapted the UTAUT Questionnaire from [75, 79]. We collected the following constructs that could be correlated with the BI of using an eXplainable AI system: Performance Expectancy, Effort Expectancy, Attitude Towards Technology, Social Influence, Facilitating Conditions, Image, Relevance, Output Quality, Result Demonstrability (5-point Likert scale, from 1="strongly disagree to

5="strongly agree"). Finally, we measured the perceived explanation quality using the explanation satisfaction scale (5-point Likert scale, from 1="strongly disagree to 5="strongly agree") proposed in [41] and collected qualitative feedback using open-ended questions on participants' experience using the two AI interfaces. We now proceed to illustrate the two AI interfaces used in our experiment.

3.4 Interface Dr.AI: Only Suggestion.

Acting as clinical DSS, we used Doctor AI [18], a Recurrent Neural Network able to predict patients' future diagnoses based on their past clinical histories. We post-processed Doctor AI outcomes transforming them from multi-label (every diagnosis of future visits) to binary to predict whether a patient would have an acute MI or not. A static visualization of the interface providing only Doctor AI suggestions is shown in figure 1. The visits of the patients are represented as a set of grey dots, and each dot represents a condition diagnosed in the corresponding visit. For example, this patient was diagnosed with five conditions in their first visit and three conditions in the second one. In the dynamic visualization, participants were able to explore the conditions diagnosed in each visit and visualize their descriptions by moving the cursor over the corresponding dots. Finally, the AI suggestion is shown in red to the left of the patient's clinical history.

3.5 Interface Dr.XAI: Suggestion and Explanation.

To extract an explanation for the algorithmic suggestion, we employed Doctor XAI [66], an eXplainable AI (XAI) technique able to deal with sequential clinical histories that use medical knowledge in its explanation extraction process. Doctor XAI's explanations highlight which conditions in the clinical history of the patients were deemed most important by the algorithm in its decision-making process. Furthermore, Doctor XAI also provides information regarding the missing conditions that influenced the algorithmic decision. A static visualization of the interface providing AI suggestions and explanations is shown in figure 2. Doctor XAI assigns a different color to each dot according to the corresponding condition's relevance to the algorithmic decisions. Dots corresponding to conditions deemed irrelevant are left grey, while dots deemed relevant are colored blue. Furthermore, Doctor XAI shows as yellow dots conditions that are missing from the patient's clinical history that would have changed algorithmic suggestion. Finally, a summary of the explanation is written under the algorithmic suggestion. The dynamic visualization allowed participants to highlight the conditions in the clinical history corresponding to each sentence in the written explanation summary. Dr.XAI's explanations are both medical domain-aware and a good representative of a common type of AI explanation: the *removal-based type* of explanation [21]. Like other popular removal-based approaches, Dr.XAI explanations summarize each feature's influence on the model outcome [51, 68]. However, unlike other removal-based approaches, it also employs medical knowledge in the explanation extraction process, meaning that the features highlighted to be important were selected considering their medical meaning. These explanation characteristics are well suited for our purpose of evaluating the impact of AI explanations on healthcare providers.

4 RESULTS

4.1 Quantitative analysis

A total of 31 healthcare providers participated in the online experiment. The analysis discarded three participants: one did not pass the attention check question, while two gave 100 as their initial estimate, which yielded undefined values for the WOA ($A = I$). Eventually, 28 participants were retained for the study. 5 doctors, 20 nurses, one health care assistant, one dietetic assistant practitioner, and one ambulance call dispatcher. The mean age was 41 years old (SD=11) ranging from 24 to 65 years old. 21 were women and 7 men. The male sample has a mean age of 34 years old (SD=9), and the female sample has mean 43 years old (SD=11). We performed all the analysis in Python.

Weight of Advice and Confidence. In figure 3(a), we show the result of the comparison between the WOA for the two AI interfaces: Dr.AI (only suggestion) and Dr.XAI (suggestion and explanation). The WOA was higher for the Dr.XAI interface (Mdn=0.31) than the Dr.AI interface (Mdn=0). A paired-samples two-sided Wilcoxon signed-rank test indicated that this difference was statistically significant ($T = 32.5$, $p = 0.002$). This result confirmed our first hypothesis, showing that participants were more influenced by the AI interface explaining its recommendation. Since advice-taking is positively correlated with trust, we can interpret this result by saying that, on average, participants implicitly trusted more the AI interface that provides explanations. In figure 3(b) we compared participants' confidence shift for the two interfaces. The confidence shift was measured as the difference of the reported participant's confidence in the estimate before and after receiving the AI advice. A paired-samples two-sided Wilcoxon signed-rank test did not find any statistically significant difference between the two interfaces $T = 169$, $p = 0.869$. This means that the explanation did not significantly increases or decreased participants confidence in their second estimate compared with a system that provide only the suggestion.

Behavioral Intention and Explicit Trust In figure 3(c) we compared the behavioral intention of use for the two AI interfaces. A paired-samples two-sided Wilcoxon signed-rank test did not find any statically significant difference between the two interfaces $T = 37$, $p = 0.076$. This result did not confirm our second hypothesis that the behavioral intention of using the AI interface Dr.XAI (suggestion and explanation) was higher than the Dr.AI (only suggestion) one. However, our results also indicated a significant positive Spearman correlation between the behavioral intention of using the Dr.XAI interface and the perceived explanation quality $rs(27) = 0.67$, $p < .001$. Similarly, we did not find a significant difference in explicit trust between the two interfaces (figure 3(d), paired-samples two-sided Wilcoxon signed-rank test, $T = 157.0$, $p = 0.881$), but we found a strong positive Spearman correlation between explicit trust and perceived explanation quality ($rs(27) = 0.77$, $p < .001$). This could indicate that this particular type of explanation does not suit healthcare providers well. Indeed, like those of most state-of-the-art XAI methods, such an explanation was developed and designed with debugging purposes in mind rather than to fit the specific needs of the final user. Therefore, healthcare providers perceive this explanation as unsatisfactory

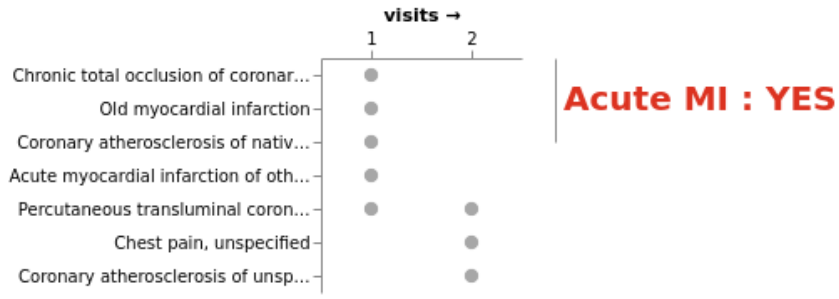

Figure 1: Static visualization of the *only suggestion* AI interface

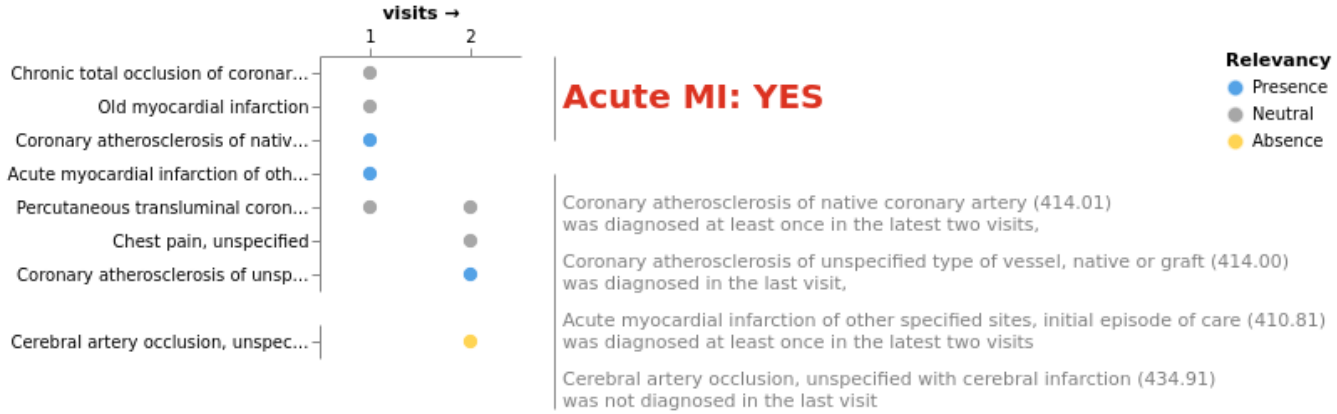

Figure 2: Static visualization of the *suggestion and explanation* AI interface

and do not increase their behavioral intention of use or trust in the system when presented with it.

Further Findings Table 1 a comparison between the UTAUT variables in the two interfaces together with their medians and the related paired-sample Wilcoxon signed-rank test statistics and its p-value. Following Bonferroni's correction method, a more stringent alpha (α=.005) was set for these particular tests. According to our correction, no significant differences have been found in the dimensions of acceptance between the two systems. Given the small sample size, we leave to future works the creation of two models investigating which factors impact the most the behavioral intention. Furthermore, no statistically significant correlation between the confounding variables, the WOA, and the behavioral intention was found with Spearman correlation tests. The only relevant negative correlation was found between the WOA of the Dr.AI interface (only suggestion) and the single-item measure of familiarity with the task (rs(27)=-0.58, p-value = 0.001). This means that the algorithmic suggestion had a stronger influence on participants less familiar with estimating the chances of an acute MI. Finally, a Wilcoxon signed-rank test showed a slight difference in the WOA between the different types of healthcare providers $T = 2.56$, $p = 0.025$. However, given the small sample for each category, we leave such an analysis for further works.

4.2 Open-ended Questions Insights

In order to evaluate participants' impressions, we asked them to answer open-ended questions. Participants' open-ended responses were coded through thematic coding [67]. Specifically, the analysis was carried out to create as few categories as possible without making them too broad.

Participants' perceptions and preferences Understanding users' preferences for one interface over the other is of pivotal importance to analyze their impressions. We asked the participants to give us answers about: 1)their general impression of each interface, 2) what they liked the most about the interface they had just used, 3)what they disliked the most about the interface they just used. Most participants appreciated the two interfaces, with slightly more participants leaving positive comments on the Dr.XAI interface (Dr.AI= 39.29%; Dr.XAI= 53.57%). Indeed, most participants did not appreciate the simple suggestion provided by the Dr.AI interface without any other information (54% of the participants asked for an explanation, while 46% did not express any opinion):

> *It is simple. Too simple in fact.* F, 36, Nurse
>
> *I wish this AI interface would provide more information about how it reached it's decision.* F, 40, Nurse.

However, when provided with the explanation, they were left unsatisfied by it:

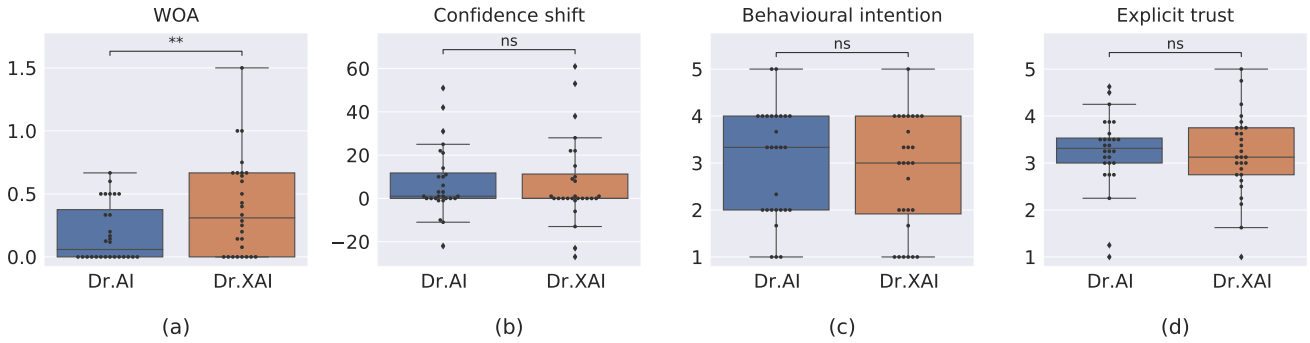

Figure 3: Boxplot comparing the WOA (a) the confidence shift after the advice (b) the behavioral intention of use and (c) the explicit trust in the two systems (d).

UTAUT variable	median Dr.AI	median Dr.XAI	Wilcoxon statistic	p-value
Performance Expectancy	3.2	3.0	66.0	0.391
Effort Expectancy	3.6	3.5	66.5	0.016
Social Influence	3.2	3.5	74.5	0.403
Facilitating Conditions	3.2	3.5	79.5	0.333
Attitude toward techology use	3.2	3.1	100.0	0.587
Image	2.0	2.2	31.5	0.325
Relevance	3.7	3.3	40.5	0.726
Output quality	3.2	3.0	64.0	0.553
Result Demonstrability	3.8	3.8	128.0	0.224

Table 1: Comparison of UTAUT variables for the two interfaces. Median, paired sample Wilcoxon signed-rank test statistics and p-value.

> *Using the AI interface with the explanation built in was something I anticipated making the decision easier, but in fact this was not the case. All the information presented too much on the screen and took a lot of time to interpret and synthesise. Decision-making became more of a lengthy and arduous process.* F, 24, Doctor.

> *I think it has a lot of potential, but would like a more detailed rationale of why it thinks an MI is likely and a numeric assessment of how likely (as I was asked to give).* F, 51, Doctor.

Some suggested implementing a natural language version of the explanation and adding the time between visits. Overall, participants did not encountered many difficulties (Dr.AI =85 %; Dr.XAI=68%). One of the common issues was understanding how to interact with the explanation. The explanation interface was considered useful to prevent novices from making mistakes and during collaborative decision-making tasks:

> *It would prevent novices making mistakes.* F, 52, Doctor.

> *The doctors in our acute medical department are very keen to discharge patients home; leaving nurses in a difficult predicament when we don't agree with their decision making. A tool such as this, could help nurses to justify their reasons for keeping a patient in hospital or to use cardiac monitoring vs. not monitoring.* F, 36, Nurse.

Algorithm aversion and fear of being replaced Eventually, one of the most surprising findings we came across is related to the participants' perceived threat of being replaced by the AI system. In both conditions, comments like the ones reported below were common:

> *Can be useful but does not replace human judgement.* F, 59, Nurse. (Dr.XAI condition).

> *it could be taken as fact that the AI is correct which disregards the human factor and individuality.* F, 53, Nurse. (Dr.AI condition)

> *It was really good but human health isn't always black and white. You can't put AI in human nature. Yes it may use stats probabilities etc but there's always that one patient that goes against the rules. I'd use it to as a tool to bear in mind but I wouldn't rely on it. [...] It takes away the thinking this the prestige of all the effort and study you've put in!.* F, 39, Nurse (Dr.XAI condition).

While this might be associated with the phenomenon of *algorithm aversion* [22], or the human discount of algorithmic advice [50], the prevailing sentiment emerging from such open-ended questions was the fear of being replaced by AI. This fear of being replaced is often an underestimated aspect in computer science

research, however, the understanding of the sociocultural environment in which the user operates has a paramount relevance in the acceptance of such AI systems [28].

5 DISCUSSION, CONCLUSIONS AND FUTURE WORK

In this paper, we presented the results of an online user study on the impact of receiving an explanation for an algorithmic suggestion in the healthcare context. In particular, we adopted the specific lens of the Weight of Advice (WOA), the Trust Scale, and the Behavioral Intention from the TAM model. We compared two interfaces for an AI-based clinical DSS by manipulating how the suggestion was presented to the healthcare providers (with or without explanation) and asked them to perform the estimation task before and after interacting with the two interfaces. We found that participants were keener on taking advice from the AI interface that explained its suggestion than the one that did not. This was reflected in a greater shift in the estimates provided after receiving such algorithmic advice, i.e., the weight of advice. We gain even more insight on the effect of the explanation on advice-taking from the open-ended questions. The answers suggested that participants did not appreciate the suggestion alone and preferred an explanation for it. However, the explanation provided left most of them unsatisfied. It is interesting to notice that, despite the low perceived explanation quality, participants were influenced by it and relied more on the advice of the AI system. This finding might be in line with previous research on *automation bias* in medicine, i.e., the tendency to over-rely on automation [34, 40, 49], and will definitely be the subject of future works. We also studied the confidence after the advice and the explicit trust in the system, finding no significant differences between the two interfaces. Similarly, we did not find a significant difference in the behavioral intention (BI) of using the two interfaces. A possible explanation for it is the high correlation between the BI and the perceived explanation quality, i.e., the proposed explanation was not appropriate for the healthcare audience. However, from the open-ended questions emerged an alternative interpretation of this finding. Indeed, many participants showed some degree of algorithm aversion and expressed the fear of being replaced by the AI system. Indeed, the AI system was perceived as threatening human judgment rather than as a decision support tool. This finding is relevant in the design of AI applications in healthcare regardless of XAI explanations and shows that it is crucial to have an interdisciplinary approach to fully comprehend the factors that influence technology adoption. Indeed, it is important to enhance the research through ethnographic methods in a triadic approach involving the user, the automation in a specific decisional context [9].

This study has some limitations. First of all, the small sample size. In future work, we aim to carry out a more complete and accurate study differentiating different healthcare providers' needs, also considering different task-related expertise. Collecting more data would also allow us to run an analysis to create a model of the UTAUT factors that most influence the BI in the healthcare context for the two interfaces. Furthermore, this study focused only on one type of AI explanation: the removal-based type. While this is a popular kind of AI explanation, future works will be dedicated to testing the efficacy of different types of explanations. Another aspect worth exploring is the case of wrong algorithmic suggestions. Our study considered only accurate algorithmic suggestions, i.e., patients correctly predicted by the algorithm as having an acute MI. In future work, we would like to investigate the relationship between trust and algorithmic accuracy. Finally, we would like to run an in-person experiment to ensure higher participants engagement.

ACKNOWLEDGMENTS

This work was supported by the European Community Horizon 2020 programme under the funding scheme ERC-2018-ADG G.A. 834756 (*"XAI: Science and technology for the eXplanation of AI decision making"*). We thank Salvatore Rinzivillo and Daniele Fadda for the technical support.

REFERENCES

[1] European Commission 2018. *EU General Data Protection Regulation.* European Commission. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf
[2] Barbara D Adams, Lora E Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol McCann. 2003. Trust in automated systems. *Ministry of National Defence* (2003).
[3] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences* 11, 11 (2021), 5088.
[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).
[5] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence* (2021), 1–10.
[6] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.
[7] Alan F. Blackwell. 2021. Ethnographic artificial intelligence. *Interdisciplinary Science Reviews* 46, 1-2 (2021), 198–211. https://doi.org/10.1080/03080188.2020.1840226
[8] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076* (2021).
[9] Clark Borst. 2016. Shared mental models in human-machine systems. *IFAC-PapersOnLine* 49, 19 (2016), 195–200.
[10] Andrea Brennen. 2020. What Do People Really Want When They Say They Want" Explainable AI?" We Asked 60 Stakeholders.. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.
[11] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.
[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
[13] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.
[14] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307.
[15] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety science* 47, 9 (2009), 1260–1270.
[16] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[17] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. *Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[18] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference.* PMLR, 301–318.

[19] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 787–795.

[20] Giovanni Comandé. 2020. Unfolding the legal component of trustworthy AI: a must to avoid ethics washing. *Version Accepted for Annuario di Diritto Comparato e di Studi Legislativi, forthcoming* (2020).

[21] Ian Covert, Scott Lundberg, and Su-In Lee. 2020. Explaining by removing: A unified framework for model explanation. *arXiv preprint arXiv:2011.14878* (2020).

[22] Berkeley Dietvorst and Soaham Bharti. 2019. People Reject Even the Best Possible Algorithm in Uncertain Decision Domains. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3424158

[23] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31, 10 (2020), 1302–1314.

[24] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[25] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[26] Jinyun Duan, Yue Xu, and Lyn M Van Swol. 2020. Influence of self-concept clarity on advice seeking and utilisation. *Asian Journal of Social Psychology* (2020).

[27] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–19.

[28] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12424 LNCS (2020), 449–466. https://doi.org/10.1007/978-3-030-60117-1_33 arXiv:2002.01092

[29] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–6.

[30] European Parliament. 2021. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS.* https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

[31] Wenjuan Fan, Jingnan Liu, Shuwan Zhu, and Panos M Pardalos. 2018. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research* (2018), 1–26.

[32] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[33] Francesca Gino and Maurice E Schweitzer. 2008. Take this advice and shove it. In *Academy of Management Proceedings*, Vol. 2008. Academy of Management Briarcliff Manor, NY 10510, 1–5.

[34] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[35] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey Of Methods For Explaining Black Box Models. *ACM CSUR* 51, 5, Article 93 (Aug. 2018), 42 pages.

[36] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[37] Ronan Hamon, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio Sanchez. 2021. Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 549–559.

[38] Yukinori Harada, Shinichi Katsukura, Ren Kawamura, and Taro Shimizu. 2021. Effects of a Differential Diagnosis List of Artificial Intelligence on Differential Diagnoses by Physicians: An Exploratory Analysis of Data from a Randomized Controlled Study. *International Journal of Environmental Research and Public Health* 18, 11 (2021), 5562.

[39] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes* 70, 2 (1997), 117–133.

[40] Steven D Hillson, Donald P Connelly, and Yuanli Liu. 1995. The effects of computer-assisted electrocardiographic interpretation on physicians' diagnostic decisions. *Medical Decision Making* 15, 2 (1995), 107–112.

[41] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[42] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[43] C Krittanawong. 2018. The rise of artificial intelligence and the uncertain future for physicians. *European journal of internal medicine* 48 (2018), e13–e14.

[44] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations *(AIES '20).* Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[45] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial intelligence in medicine* 94 (2019), 42–53.

[46] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[47] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–13.

[48] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–15.

[49] Thomas Lindow, Josefine Kron, Hans Thulesius, Erik Ljungström, and Olle Pahlm. 2019. Erroneous computer-based interpretations of atrial fibrillation and atrial flutter in a Swedish primary health care setting. *Scandinavian journal of primary health care* 37, 4 (2019), 426–433.

[50] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[51] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems.* 4768–4777.

[52] Gianclaudio Malgieri and Giovanni Comandé. 2017. Why a right to legibility of automated decision-making exists in the general data protection regulation. *International Data Privacy Law* (2017).

[53] Vidushi Marda and Shivangi Narayan. 2021. On the importance of ethnographic methods in AI research. *Nature Machine Intelligence* 3, 3 (2021), 187–189.

[54] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. 2021. Exemplars and Counterexemplars Explanations for Image Classifiers, Targeting Skin Lesion Labeling. In *2021 IEEE Symposium on Computers and Communications (ISCC).* IEEE, 1–7.

[55] Martijn Millecamp, Sidra Naveed, Katrien Verbert, and Jürgen Ziegler. [n.d.]. *To Explain or Not to Explain: the Effects of Personal Characteristics When Explaining Feature-based Recommendations in Different Domains.* Technical Report.

[56] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[57] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 417–431.

[58] Jessica Morley, Caio CV Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. 2020. The ethics of AI in health care: A mapping review. *Social Science & Medicine* (2020), 113172.

[59] Henrik Mucha, Sebastian Robert, Ruediger Breitschwerdt, and Michael Fellmann. 2021. Interfaces for Explanations in Human-AI Interaction: Proposing a Design Evaluation Approach. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–6.

[60] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.

[61] Emanuele Neri, Francesca Coppola, Vittorio Miele, Corrado Bibbolino, and Roberto Grassi. 2020. Artificial intelligence: Who is responsible for the diagnosis?

[62] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[63] Cecilia Panigutti, Riccardo Guidotti, Anna Monreale, and Dino Pedreschi. 2019. Explaining multi-label black-box classifiers for health applications. In *International Workshop on Health Intelligence.* Springer, 97–110.

[64] Cecilia Panigutti, Anna Monreale, Giovanni Comandé, and Dino Pedreschi. 2022. Ethical, societal and legal issues in deep learning for healthcare. In *Deep Learning in Biology and Medicine*. World Scientific Publishing.

[65] Cecilia Panigutti, Alan Perotti, André Panisson, Paolo Bajardi, and Dino Pedreschi. 2021. FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management* 58, 5 (2021), 102657.

[66] Cecilia Panigutti, Alan Perotti, and Dino Pedreschi. 2020. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 629–639.

[67] Michael G Pratt. 2009. From the editors: For the lack of a boilerplate: Tips on writing up (and reviewing) qualitative research.

[68] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[69] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 431–449.

[70] Jessica M Schwartz, Amanda J Moy, Sarah C Rossetti, Noémie Elhadad, and Kenrick D Cato. 2021. Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review. *Journal of the American Medical Informatics Association* 28, 3 (2021), 653–663.

[71] Lucy Shinners, Christina Aggar, Sandra Grace, and Stuart Smith. 2020. Exploring healthcare professionals' understanding and experiences of artificial intelligence technology use in the delivery of healthcare: an integrative review. *Health informatics journal* 26, 2 (2020), 1225–1236.

[72] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.

[73] Janet A Sniezek and Timothy Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational behavior and human decision processes* 62, 2 (1995), 159–174.

[74] Janet A Sniezek and Lyn M Van Swol. 2001. Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes* 84, 2 (2001), 288–307.

[75] MT Spil and WR Schuring. 2006. E-Health Systems Diffusion and Use: The Innovation. *The Users and the Use IT Model* (2006).

[76] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter PC Boon, and Ellen HM Moors. 2020. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology* 30 (2020), 5525–5532.

[77] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference*. PMLR, 359–380.

[78] Eric Topol. 2019. *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

[79] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.

[80] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[81] Himanshu Verma, Roger Schaer, Julien Reichenbach, Jreige Mario, John O Prior, Florian Evéquoz, and Adrien Raphaël Depeursinge. 2021. On Improving Physicians' Trust in AI: Qualitative Inquiry with Imaging Experts in the Oncological Domain. (2021).

[82] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, 318–328.

[83] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[84] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. 223–227.

[85] Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical Attention Propagation for Healthcare Representation Learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 249–256.

[86] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. 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*. 295–305.