Explainable recommendations and calibrated trust: two systematic user errors

Mohammad Naiseh
Department of Electronics and Computer Science, University of Southampton, UK

Deniz Cemiloglu
Faculty of Science and Technology, Bournemouth University, UK

Dena Al-Thani
College of Science and Engineering, Hamad Bin Khalifa University, Qatar

Nan Jiang
Faculty of Science and Technology, Bournemouth University, UK

Raian Ali
College of Science and Engineering, Hamad Bin Khalifa University, Qatar

***Abstract*— The increased adoption of collaborative Human-AI decision-making tools triggered a need to explain the recommendations for safe and effective collaboration. However, evidence from the recent literature showed that current implementation of AI explanations is failing to achieve adequate trust calibration. Such failure has lead decision-makers to either end-up with over-trust, e.g., people follow incorrect recommendations or under-trust, they reject a correct recommendation. In this paper, we explore how users interact with explanations and why trust calibration errors occur. We take clinical decision-support systems as a case study. Our empirical investigation is based on think-aloud protocol and observations, supported by scenarios and decision-making exercise utilizing a set of explainable recommendations interfaces. Our study involved 16 participants from medical domain who use clinical decision support systems frequently. Our findings showed that participants had two systematic errors while interacting with the explanations either by skipping them or misapplying them in their task.**

■ **THE INTRODUCTION** Current advances in machine learning have increased the enactment of human-AI collaborative decision-making tools in safety-critical applications such as medical systems and military applications [6]. Researchers have identified trust calibration as the main requirement for safe and responsible implementation for such tools in everyday scenarios [1,2]. Trust calibration is the process of successful judgment of the main components of trust: cognition-based trust and affect-based trust [2,3]. Trust is calibrated when the human operator can understand and adjust their level of trust to the current state of the AI [3]. This adjustment is crucial due to the dynamic and uncertain nature of AI-based applications. When users fail to manage their trust, they either end-up with over-trust, e.g., people follow incorrect recommendations or under-trust, and they reject a correct recommendation. Previous research [3] identified five primary contexts where trust calibration errors in automation occur, their reasons for occurrences and potential design solutions. Overall, trust calibration errors can happen when users do not understand the system functionality, do not know its capability, overwhelmed with the system output, lack situation awareness or feel a loss of control the system. Such faulty in design has shown critical safety issues [3].

Research in eXplainable AI (XAI) showed that augmenting AI-based recommendations by explanations can enhance trust calibration as it can give human decision-makers insights and transparency on how the AI arrived at its recommendation. Explanations are supposed to support users in developing correct mental models of the AI, identifying situations when recommendations are correct or incorrect, and mitigating trust calibration errors [1, 2, 4]. However, recent evidence suggests that explainable AI-based systems also have not improved a successful trust calibration as users' still, on average, end-up in situations where they over-trust or under-trust the AI-based recommendations [2,21]. In the context of XAI and trust calibration, previous work has typically focused on evaluating explanations in trust calibration context [21] and identifying explanations types [2] and presentation formats [23] for improved trust calibration. In general, the work often assumed that people would engage cognitively with each explanation and use its content to build a correct mental model and improve trust calibration. However, this assumption can be incorrect; humans often reluctant to engage in what they perceived as effortful behavior [24] resulting in less informed trust decisions.

Indeed, some studies demonstrated situations where explanation failed to enhance users' trust calibration, e.g., explanations were perceived as an information overload [1]. Others also related the failure of explanation to improve trust calibration errors to human behavior and cognitive biases, e.g., cognitive laziness of humans to read explanations [22]. Despite the emerging need to design effective XAI interfaces to calibrate users' trust, there is a need for more knowledge about situations and contexts in which explanations do not enable adequate trust calibration, i.e., what kind of scenarios or errors could happen in real-time.

To this end, we aim to explore people interaction behavior with explanations in Human-AI collaborative decision-making tasks. Such a knowledge would ultimately inform future design affordances and aid researchers and designers in developing effective calibrated trust XAI interfaces. In this study, we pose the following research questions:

- *How do users interact with explanations during their Human-AI collaborative decision-making task?*
- *What are those situations where users fail to calibrate their trust in the presence of explanations?*

To answer these questions, we conducted a two-stage qualitative study which involved 16 participants (doctors and pharmacists) who use AI-based decision-support tools frequently in their clinical settings. Our results include a qualitative investigation of people interaction behavior with AI explanations that revealed two systematic users' errors, leading to trust calibration flaws and their reasons.

2. RESEARCH METHOD

We conducted a think-aloud protocol where participants were asked to perform Human-AI collaborative decision-making task. We then conducted follow-up interviews to gain more insights and discuss our observation on participants' experience during the task. To help our investigation, we designed an AI-based decision-support mock-up tool that is meant to support medical practitioners in classifying the prescriptions into confirmed or rejected. Prescription classification is a process that medical experts in a clinic follow to ensure that a prescription is prescribed for its clinical purpose and fit the patient profile and history. We designed the mock-up based on template and interfaces that are familiar to our participants in their everyday decision-making tasks. The scenarios simulated a diversity of conditions and explanation types that the decision-maker could face in the real-world scenarios where trust calibration errors could happen, e.g. imperfect AI recommendation due to the dynamic nature of the application. Hence, we included both correct and incorrect recommendations of each class. We chose prescription classification case study as it reflects a high-cost decision-making task performed collaboratively between the human expert and the AI. In [7], we explain more about the research method and material used.

2.1. RECRUITMENT AND PARTICIPANTS

We approached three hospitals in the UK by sending an email invitation and got a positive response from 16 individuals. No more participants were approached due to the fact that during the data analysis, resulted themes and codes became eventually repetitive. We followed the principles of reaching the saturation point in qualitative methods in [5]. This was a reasonable assurance that further data collection would introduce similar results and would confirm the existing themes. Details about the population are provided in Table 1. A study protocol was developed, and pilot tested with two practitioners, one medical academic and one AI expert.

Table 1. Population details

Variable	Value	N=16	%
Age	20-30 30-40 40-50	5 7 4	31.25% 43.75% 25%
Gender	Male Female	10 6	62.5% 37.5%
Role	Doctors Pharmacists	4 12	25% 75%
Experience	<5 5-10 10-15 >15	4 8 3 1	25% 50% 18.75% 6.25%
Hospital	A B C	6 6 4	37.5% 37.5% 25%

2.2. CONSENT PROCEDURE

First, the participants were briefed about the study, verbally and through a written participant information sheet. They were then asked to sign a contest form. Participants were also asked a number of questions about themselves, such as their experience. For enhancing the validity of the collected data, we designed the study to avoid promoting participants to think about explanations and trust calibration as a main objective of the study. We initially demonstrated the study purpose, describing it as an investigation on how medical practitioners use AI-based tools in their work environment. We also mentioned that AI-based tool can explain why a recommendation has been made. Participants were told they could discontinue the study at any point. We debriefed the participants after the study about the detailed purpose of the study.

2.3. STUDY PROCEDURE

We gave each of our participants ten scenarios that included AI-based recommendations. Each scenario was accompanied by an explanation. We used five explanation types revealed from a recent literature review [6]: Local, Global, Example-based, Counterfactual and Confidence explanations. The scenarios presented to our participants were hypothetical scenarios designed with collaboration with a medical oncologist. We designed the scenarios to be clear, challenging and not trivial so that recommendations, explanations and trust calibration were all substantial processes. This ultimately helped to put our participants in a realistic setting: exposing to an AI-based recommendation and its explanations where trust calibration is needed and where errors in that process are possible. The 16 participants were asked to make decisions considering the patient profile, the recommendation and the explanations and whether to follow the AI-based recommendation if they see it as correct or reject it if they see it as incorrect. For each scenario, participants were encouraged to think aloud during their decision-making process. They were asked to think freely and encouraged to make optimal decisions. Each of the participants completed ten scenarios representing two cases (correct and incorrect) of each of the five explanation classes. This resulted in 160 completed decision-making tasks. The researcher observed, audio-recorded the sessions and took notes. Finally, we invited our participants to a follow-up interview about their task and explainability experience. Figure 1 summarizes the study workflow.

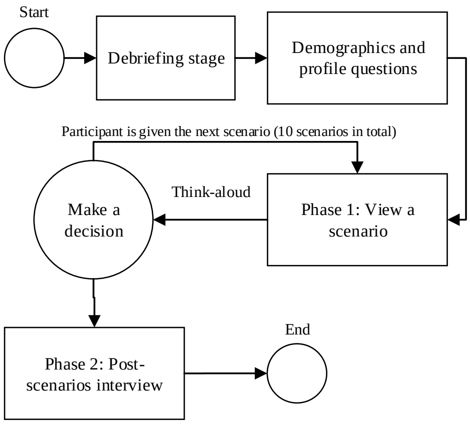
Start
Debriefing stage
Demographics and profile questions
Participant is given the next scenario (10 scenarios in total)
Think-aloud
Make a decision
Phase 1: View a scenario
End
Phase 2: Post-scenarios interview

Figure 1 Study workflow for each participant

2.4. DATA ANALYSIS

Two sets of data were collected and used to answer our research questions in this study. The first consisted of the transcript of audio files of both of the study stages (the think-aloud and the follow-up interviews). The second is the researchers' notes, which contained their observations of participants' behavior and interaction style with the XAI interface. For qualitative data, we performed a content analysis with the Nvivo tool's support. The authors had an initial meeting where they agreed a common grounds and analysis scope and style. The analysis was mainly done by the first author. The analysis was reviewed iteratively by the others through frequent meetings which led to split, modify, discard or add categories to ensure that all responses and their contexts were well represented and categorized.

2.5. STRENGTHS AND LIMITATIONS

Scenarios in combination with the think-aloud approach has been shown to be valuable for gaining insight into decision-making mechanisms [8]. An additional strength of this study was the variety of explanation types used in scenarios, which triggered different responses from participants. All participants were shown the same ten scenarios. Since our sample included three different hospitals in the UK, the results are not limited to a specific practice. Furthermore, participants differed in experience, age and gender, making the sample diverse within this specific field of expertise.

Although scenarios were created to reflect daily practice, practitioners often emphasized additional steps that they would normally take before reaching a decision, such as discussing with colleagues and meeting with patients. These options were not available in the study in which practitioners could only express their desire to know more information about the scenarios. This caused practitioners to work with their own knowledge and the available explanations instead of offering them the possibility to investigate their uncertainties. Furthermore, the think-aloud methodology does not ensure that all thoughts behind a decision are explicit. Some decision-making steps might have been applied implicitly, i.e. as a tacit knowledge [9] which might have been the case for user's interaction behavior with explanations. We tried to mitigate that through follow-up interviews. Finally, study is qualitative involving a relatively small sample and our results are yet to be tested for generalizability. Our main purpose is to shed light on important design considerations when designing XAI for calibrated trust goal. Longitudinal studies and more objective measures, possibly through experimental design, are still needed to validate our results and map them to explanation types.

3. RESULTS

In this section, we report on our studies' results that are related to systematic users' errors when interacting with explanations. Through observations and think-aloud, we investigate reasons why explanations may not improve trust calibration focusing on the cognitive dimension of trust within a sample of professionals in the medical domain. Our results indicated that for this trust facet and sample, users' errors were the main source of errors in trust calibration leading to making an incorrect decision. However, these users' errors may not be exclusive to calibrated trust design goal. In addition, such errors could also be linked to other design goals for XAI interfaces, such as perceived fairness of the AI [25] and explanation usability [26].

Our analysis showed three main themes of users' behavior: *skipping, applying* and *misapplying*. Within the scope of this paper, we only focus on *skipping and misapplying* themes that relate to errors in trust calibration in the presence of explanations. We considered an error as systematic if it happened for all explanation types. We also required that these errors crosscut all scenarios to avoid a case where an issue stems from one or a few scenarios and designs. Figure 2 shows a frequency analysis of behaviors when interacting with 160 interactions with explanation interfaces and the emerged themes (ten were shown to each of 16 participants).

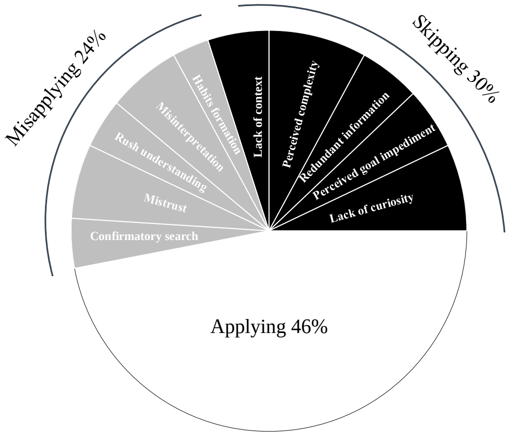

Figure 2. Participants' interaction behavior with explanations.

3.1. SKIPPING EXPLANATIONS

Explanations might fail to support trust calibration process when they are skipped. We observed that some participants made decisions collaboratively with our AI-based decision-making tool without thoroughly reading explanations. In the following sections, we describe the main reasons for errors in the Skipping category.

Lack of curiosity. Curiosity describes the desire to know, learn, or experience an explanation [15]. During the study, participants showed a lack of curiosity to seek an explanation from the AI-based tool. Participants did not feel that the explanation motivated them to learn new ideas, resolve knowledge and solve problems. P5 mentioned, *"... to be honest with you, I was not really interested in reading the explanation ... I mean I did not feel that could add something new to me*". Previous research showed that humans are selective when being curios to seek for explanation and depend on the context and individual characteristics [10]. For example, people might be more curious to read an explanation when the recommendation does not meet their expectations. Furthermore, in scenarios when the explanation contained too many features and information, participants' degree of curiosity was low, and participants were silent during these scenarios. Such situations led participants to skip explanations and discourage them from engaging in what they perceived as effortful processing behavior.

Perceived goal impediment. Participants skipped the explanations that they perceived as a goal impediment. During the study, several participants were focused on finishing the task and making decisions with an AI-based tool rather than reading the explanation. According to reversal theory [11]; individuals in a serious-minded state have a high goal orientation, while those in the playful-minded state have low goal orientation. People in high critical decision-making environments are likely to be in a serious-minded state, where additional information might be prone to be perceived as a goal impediment. Furthermore, perceived goal impediment could be relating to factors such as time constraint and multi-tasking. P12 mentioned, *"... that [explainability] experience was good in general ... but I doubt that it could work in real-world ... doctors and pharmacists are too busy to validate each decision with an explanation*". Similarly, P6 added, *"... I cannot see how these explanations will work in everyday prescriptions screening*". Such interruption into users' tasks leads to psychological reactance and results in users' avoidance [12]. Previous research used the theory of psychological reactance to explain users' avoidance of online advertisement content [12]. This theory shows that people tend to be psychologically aroused when they perceive their freedom to be threatened by others. This tendency leads individuals to restore threatened freedom by reacting to the threat. In the field of communication, the theory of psychological reactance offers an explanation for why persuasive messages, including explanations, can sometimes produce odds with their intent. Humans reject or move away from a message if the message threatens or attempts to reduce his or her personal freedom of the decision. We argue that increasing users' perceived value of the explanations would make them less likely to be skipped. For instance, the explanation design might bind into regret aversion bias [13], e.g. people might become more careful in reading explanations when they are informed about a certain level of risk from skipping them.

Redundant information is another cause of skipping as participants mentioned that in certain scenarios, explanations contained information that is simplistic and common sense for them. For instance, P9 stated, "*The average pharmacist does not need to see all these factors that the AI is considering, some of them are just simple rules*". Also, P6 criticized Counterfactual explanation and stated, "*... mentioning the AI could change its decision if age was 29 does not consider as* *a useful explanation in our setting ... I mean we all know that ... explanations should be smart enough*". Research in cognitive science and explanations showed that people tend to avoid circular and redundant explanations [14]. For example, people refuse an explanation such as "*this diet plan works because it helps people lose weight*". Such repetition of facts and no additional substantiation would make users lose their trust in explanations and even avoid further explanations. In general, people evaluate the meaningfulness of the explanations based on three main dimensions: Circularity, Relevance and Coherence [14]. To address this issue, previous research [15] proposed the theory of mind to suggest a design solution for achieving meaningful explanations to users in explainable AI applications. The research argued that intelligent agents should keep track of what has already been explained to users and evolve explanations over time. The adoption of adaptive and personalized user interfaces [1] would also be a potential solution direction. In summary, techniques to construct a user model, either explicitly or implicitly are required in future Human-AI collaborative decision-tools to avoid repetitive explanations.

Perceived complexity. Participants ignored explanations because they thought it would take too much time to understand them, e.g. long explanations. In contrast, shorter explanations such as Counterfactual explanation caught participants' attention. For instance, P11 ignored a Global explanation but read and engaged with Counterfactual explanation, and mentioned: "*It could be useful, but I won't bother digging what does that mean*". Participants discussed making quick judgments whether they would interact with explanations or completely skip them based on explanation length. For instance, P12 stated during Global explanation scenario, "*I would usually look for the first three or four values*". Explanations variables such as their size, number of chucks and lines showed to confuse users and made explanations less acceptable [1]. Such long explanations require more processing time and contribute to lower user satisfaction. Another factor that contributes to avoiding long explanations may involve the order in which people receive the explanations [17]. People tend to rely on information presented at the beginning when they try to form an intention to read [16]. Therefore, the order of the explanation chunks could be crucial to engage users with explanations and avoid skipping long explanations.

Lack of context. Participants ignored explanations that they could not contextualize to their everyday decision-making tasks. We found that participants were often expecting explanations to be task-centered and reflective to their domain knowledge and terminology. In a Counterfactual explanation scenario, P8 stated, "*I find this irrational, the explanation is saying the prescription would have been prescribed if the patient age is 50 ... I mean patient age is not something we can change ... I expected something like blood test or any other variable that we can do something about it*". Another case of skipping explanations due to lack of context when participants asked for additional contextual information in order to contextualize the explanation to their medical practice. P9, who skipped a Global explanation mentioned, "*I would like also to see correlations between patient information to judge whether this is valid information in this case*". Overall, participants were more motivated to engage with explanations that are reflective to their task characteristics rather than understanding the reasoning of the AI. User-centered iterative design with collaboration with domain experts, e.g. medical doctors, to identify task-centered explanations is needed.

3.2. MISAPPLYING EXPLANATIONS

Even when participant engaged with explanations and paid attention to them, we observed that they also misapplied them in their task. In the following sections, we discuss main situations that led to misapplication errors.

Misinterpretation. Some participants misinterpreted our presented explanations and that led to incorrect conclusions about explanations and recommendations. For example, P2, who is a pharmacist, mentioned that the AI-based tool is biased based on his interpretation of the Global explanation. The explanation in those scenarios gave a high importance value for patients' blood test in the recommendation. P2 stated: *"... so shall we screen all prescriptions only on blood results?".* Such misinterpretation led to distrust in the AI-based tool. Similarly, P9 had a false interpretation of a Confidence explanation and stated that "*44% certainty in a diagnosis is a good value*". Participants depended on their previous knowledge to interpret the available explanations, which led to building a wrong conclusion. It may be useful to accompany the AI-based tool with an onboarding feature that allows users to understand and familiarize themselves with explanations and their interpretations. Such a technique has been used in the literature of Human-AI interaction by Cai et al. [4] to familiarize medical practitioners with AI-based cancer prediction tool. This offered a way to aid users' in building correct mental models regarding the actual capabilities and limitations of the tool. For example, videos tutorials or FAQs could serve that goal.

Mistrust. Although our participants often assume that explanations are cooperative, they were also well prepared to mistrust them. Some participants felt that explanations were deceptive or untrustworthy to follow. Participants quickly assessed that explanation and voiced skepticism about the correctness and validity explanations. P8 noted, "*I am wondering if an experienced pharmacist has looked at this before*". Sometimes skepticism about the explanation content was combined with skepticism about the source of the explanation. For example, P5 wondered if Local explanation considered data coming from different hospitals, "*we have got to know which hospital this explanation covers, this could completely change my opinion about this explanation*". Our participant required several meta-information about the explanation to judge its correctness and solve mistrust issues. People might mistrust an explanation based on what they know about the motivations and abilities it sources [16]. Given the well-known phenomena in the psychology literature, addressing such suspicion in the XAI interface can be detrimental for user mistrust correction.

Confirmatory search. Participants did not read the full explanation and searched for information that confirmed their initial hypothesis, i.e., they were selective in what to read and rely on. When shown an Example-based explanation, P4 who is a pharmacist stated, "*Well, I would look for the examples that I've already experienced in the past*". During the study, participants did not take into consideration disconfirming their hypothesis to correct their mental model but found confirming evidence to further strengthen their hypothesis. They completed their explanation analysis with overconfidence of their initial insights and ended up with trust calibration errors. Several variables can facilitate confirmatory search tendency during the decision-making, such as the increased number of the available information, sequential information presentation, or negative mood [17]. XAI research is to look for design techniques that encourage them to read the full explanation and avoid bias.

Rush understanding. Participants incorrectly held a belief that they understand the explanation deeper than they actually did. This effect was obvious in the interview stage, e.g., P4 stated, "*Well in many cases I could predict how the AI work after reading the explanations in first two cases*". Likewise, P7 mentioned, *"... I would say that I have a confidence to tell how it worked*". However, they failed to answer our follow-up questions that delved into the details and conclusions. Such miscalibration of their understanding is another case of the overconfidence effect [14]. Furthermore, rush understanding could also be related to the explanation itself, e.g., being incomplete or reduced, which made it difficult to have much practice in assessing ones' own understanding. One design solution could be by slowing the users down to enable the reflection over their actions.

Habits formation. As job actions and decision are typically repetitive, users collaborating with an AI-based decision-making tool are prone to develop habits [19]. During the study, participants became gradually less interested in the details of an explanation and overlooked it altogether. Such behavior is associated with the development of peoples' expectations about the behavior and the performance of the environment [12]. P4 who showed similar behavior mentioned, "*I think this is similar to the previous explanation*". Such habits could damage explanation goal to support trust calibration. For instance, doctors with a successful diagnosis experience with an AI may fail to notice a minor change in the AI accuracy and the explanation output. The continuous pairing of collaborative diagnosis with positive outcome may in time cause the act to become automatic, triggering an unconscious response which is no longer linked to the explanation output [19]. Habits might be also triggered by prior interaction in a chain of responses, by environmental cues, such as time of the day or location, or by the particular internal state such as moods [12]. XAI design is to monitor such habits formations and try to prevent it, e.g., when a user agrees excessively, an adaptative design approach can change the explanation interface structure so that it triggers a fresh thinking.

4. DISCUSSION

One main goal of communicating explanations in Human-AI collaborative decision-making is to enhance the trust calibration process. This study has examined the role of explainability in enhancing Human-AI collaborative decision-making and trust calibration process in particular. One of the key findings is that explanations failed to support users in their trust calibration process due to two primary users' errors: skipping and misapplying. We argue that building XAI interfaces that consider these errors and develop design constraints to limit them can support the explanation goal of enhancing trust calibration. For instance, we observed a high frequency of skipping explanations when participants perceived explanations as an impediment to their task. As a corollary, a design that fits the explanation in the task workflow can limit such errors and may support the trust calibration process as users would read the explanation and understand the AI reasoning.

Also, the relationship between failing to calibrate trust and user errors could be further investigated through the lens of human decision-making processes. According to the Elaboration Likelihood Model (ELM), humans process information in two different routes: a central route in which information processing is slow and reflective and a peripheral route in which information processing is fast and relies on mental shortcuts [18]. It has been suggested that individuals have the disposition to use the peripheral route as it saves time and effort, and this type of processing is especially relevant to medical settings where time constraints exist. While mental shortcuts are usually effective in decision-making, their unconscious and automatic nature make them prone to cognitive biases. Overall, implementing AI-supported decision-making tools with explanations could be a way of mitigating biases that the people might have in their everyday decision-making tasks as such explanations could activate central route processing [15]. However, human biases could also influence the processing of explanations, and this can lead decision-makers to either end-up with under-trust or over-trust. For example, under-trust may result from anchoring bias when participants look at only salient features of AI explanations and consequently judge the quality of information to be untrustworthy. Similarly, over-trust may result from confirmation bias as mentioned before when participants favor explanations that are consistent with their initial hypothesis. In this light, the presentation of explanations has the risk of further reinforcing biases that decision-makers may already have. This highlights the necessity to address cognitive biases in the design of explanations.

Finally, either skipping or misapplying explanations could be resulted from the fact that participants did not seek an explanation. Such behavior limited users' learning process of the AI reasoning and its underlying logic, so their trust was not calibrated. It has been found that despite the availability of explanations, people might utilize a small amount of them or avoid seeking explanations, even when they need them [30]. Thus, if the goal of explanations is to calibrate users' trust, effective explanation seeking behavior may contribute to improving users' learning and trust calibration processes. Our results pose a new requirement for XAI interfaces to focus, especially at the earlier stage of interacting with the AI, on increasing explanation-seeking behavior. This could be potentially implemented by applying principles of persuasive design [27] and persuasive learning [20], e.g., showing users' level of knowledge about the AI.

5. CONCLUSION

Designing explanations for trust calibration has been identified as one of the main goals for safe and effective AI-supported decision-making tools. However, it is often remaining unclear in the literature why explanations were not always supporting users' in their trust calibration. This motivated our work to explore how people interact with explanations in their Human-AI collaborative decision-making task. We focused on particular situations where explanations did not effectively support users to calibrate their trust. As a general conclusion, explainability for trust calibration might conflict with usability: trust calibration require extra efforts from the users, e.g. read and interact with the explanation. Thus, integrating explanations in Human-AI collaborative decision-making environments needs to analyze and explore the costs and benefits of favoring between explainability and usability.

ACKNOWLEDGMENTS

This work also acknowledges support from the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1). This work involved human subjects or animals in its research. Approval of all ethical and

experimental procedures and protocols was granted by Bournemouth University Ethics Committee.

■ REFERENCES

1. Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020, April. Personalising Explainable Recommendations: Literature and Conceptualisation. In World Conference on Information Systems and Technologies (pp. 518-533). Springer, Cham.
2. Zhang, Y., Liao, Q.V. and Bellamy, R.K., 2020, January. 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 (pp. 295-305).
3. Lee, J.D. and See, K.A., 2004. Trust in automation: Designing for appropriate reliance. Human factors, 46(1), pp.50-80.
4. Cai, C.J., Winter, S., Steiner, D., Wilcox, L. and Terry, M., 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), pp.1-24.
5. Faulkner, S. L., & Trotter, S. P. (2017). Data saturation. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), The international encyclopedia of communication research methods (pp. 1–2). Hoboken, NJ: John Wiley & Sons.
6. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp.82-115.
7. M Naiseh, Explainable Recommendations and Calibrated Trust – Research Protocol., 2021, Technical Report, Bournemouth University, 35306, http://eprints.bournemouth.ac.uk/35306/
8. Wolcott, M.D. and Lobczowski, N.G., 2020. Using cognitive interviews and think-aloud protocols to understand thought processes. Currents in Pharmacy Teaching and Learning.
9. Howells, J., 1996. Tacit knowledge. Technology analysis & strategic management, 8(2), pp.91-106.
10. Wrobel, G.M., Grotevant, H.D., Samek, D.R. and Korff, L.V., 2013. Adoptees' curiosity and information-seeking about birth parents in emerging adulthood: Context, motivation, and behavior. International journal of behavioral development, 37(5), pp.441-450.
11. Apter, M., 1997. Reversal theory: what is it? PSYCHOLOGIST-LEICESTER-, 10, pp.217-220.
12. Van Doorn J, Lemon K N, Mittal V, Nass S, Pick D, Pirner P, Verhoef P C. (2010). Customer Engagement Behavior: Theoretical Foundations and Research Directions. Journal of Service Research, 13(3): 253-266.
13. Caraban, A., Karapanos, E., Gonçalves, D. and Campos, P., 2019, May. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
14. Keil, F.C., 2006. Explanation and understanding. Annu. Rev. Psychol., 57, pp.227-254.
15. Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, pp.1-38.
16. Oppenheimer DM. Spontaneous discounting of availability in frequency judgment tasks. Psychol Sci. 2004; 15:100–5.
17. Fischer, P., Schulz-Hardt, S. and Frey, D., 2008. Selective exposure and information quantity: how different information quantities moderate decision makers' preference for consistent and inconsistent information. Journal of personality and social psychology, 94(2), p.231.
18. Evans, J. S. B., 2008. Dual-processing accounts of reasoning, judgment, and social cognition. Annu. Rev. Psychol., 59, 255-278.
19. Wood, W., and Rünger, D., 2016. Psychology of habit. Annual review of psychology, 67.
20. Aleven, V., Stahl, E., Schworm, S., Fischer, F. and Wallace, R., 2003. Help seeking and help design in interactive learning environments. Review of educational research, 73(3), pp.277-320.
21. Bussone, A., Stumpf, S. and O'Sullivan, D., 2015, October. The role of explanations on trust and reliance in clinical decision support systems. In 2015 international conference on healthcare informatics (pp. 160-169). IEEE.
22. Wagner, A.R. and Robinette, P., 2021. An explanation is not an excuse: Trust calibration in an age of transparent robots. In Trust in Human-Robot Interaction (pp. 197-208). Academic Press.
23. Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G. and Kaplan, L., 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns, 1(4), p.100049.
24. Wouter Kool and Matthew Botvinick. 2018. Mental labour. Nature human behaviour 2, 12 (2018), 899–908.
25. Schmitt, N., Oswald, F.L., Kim, B.H., Gillespie, M.A. and Ramsay, L.J., 2004. The impact of justice and self-serving bias explanations of the perceived fairness of different types of selection tests. International Journal of Selection and Assessment, 12(1-2), pp.160-171.
26. Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S. and Doshi-Velez, F., 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1802.00682.
27. Oinas-Kukkonen, H. and Harjumaa, M., 2009. Persuasive systems design: Key issues, process model, and system features. Communications of the Association for Information Systems, 24(1), p.28.

Mohammad Naiseh is a research fellow at the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. SO17 1BJ. His research interests include explainable artificial intelligence and human-computer interaction. Mohammad received his PhD in Explainable AI from Bournemouth University. Contact him at m.naiseh@soton.ac.uk.

Deniz Cemiloglu is a PhD candidate at the Department of Computing and Informatics, Bournemouth University, UK. Her research interests include digital addiction and responsibility by design.

Dena Al-Thani is an assistant professor at the Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Qatar. Her research interests include human-computer interaction, inclusive design, accessibility and e-health.

Nan Jiang is an Associate Professor at the Computing and Informatics, Bournemouth University, UK. His primary research field is Human Computer Interaction (HCI). He is mainly interested in developing novel usability evaluation methods for emerging technologies and needs.

Raian Ali is a Professor at the Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University, Qatar. His research focuses on the inter-relation between technology design and social requirements such as motivation, transparency, wellbeing and responsibility.