Nudging through Friction: an Approach for Calibrating Trust in Explainable AI

Mohammad Naiseh
School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
m.naiseh@soton.ac.uk

Reem S. Al-Mansoori
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
reem.almansoori.qtr@gmail.com

Dena Al-Thani
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
dalthani@hbku.edu.qa

Nan Jiang
Department of Computing and Informatics
Bournemouth University
Poole, United Kingdom
njiang@bournemouth.ac.uk

Raian Ali
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar
raali2@hbku.edu.qa

***Abstract*— Explainability has become an essential requirement for safe and effective collaborative Human-AI environments, especially when generating recommendations through black-box modality. One goal of eXplainable AI (XAI) is to help humans calibrate their trust while working with intelligent systems, i.e., avoid situations where human decision-makers over-trust the AI when it is incorrect, or under-trust the AI when it is correct. XAI, in this context, aims to help humans understand AI reasoning and decide whether to follow or reject its recommendations. However, recent studies showed that users, on average, continue to overtrust (or under-trust) AI recommendations which is an indication of XAI's failure to support trust calibration. Such a failure to aid trust calibration was due to the assumption that XAI users would cognitively engage with explanations and interpret them without bias. In this work, we hypothesize that XAI interaction design can play a role in helping users' cognitive engagement with XAI and consequently enhance trust calibration. To this end, we propose friction as a Nudge-based approach to help XAI users to calibrate their trust in AI and present the results of a preliminary study of its potential in fulfilling that role.**

Keywords— Human-AI Interaction, Explainable AI, Digital Nudging, Friction, Calibrated Trust

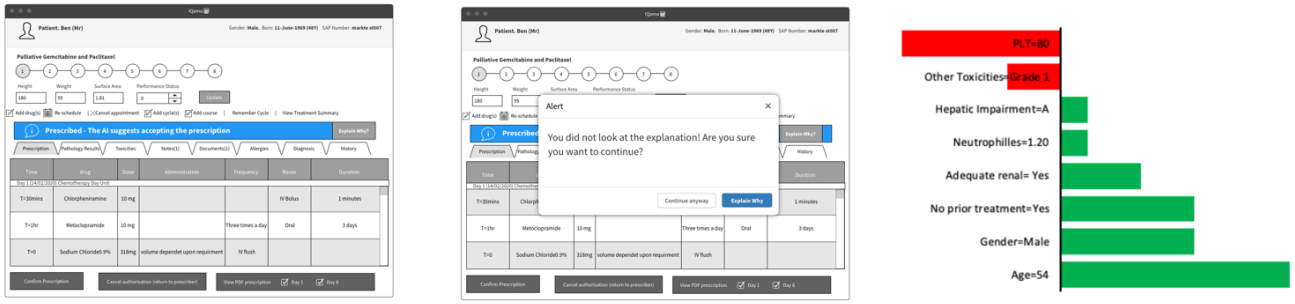
I. INTRODUCTION

Although Artificial Intelligence, specifically machine learning, has been applied in many high-stakes application domains such as healthcare, defense and justice, full automation of these applications has not yet been adopted. In many situations, human operators with domain knowledge have to ensure the accuracy and the validity of the AI outputs. While combining humans and AI in collaborative decision-making environments is expected to increase the quality of decision outcomes [1], recent studies showed that humans frequently make trust calibration mistakes and either over-trust incorrect AI recommendations or under-trust correct ones [2, 3].

Explainable AI (XAI) is a means of AI design where recommendations are supported by explanations to facilitate users' trust calibration process. Explanations provide decision-makers with insights into how the machine derived a recommendation. Explanations are supposed to help humans identify situations where AI recommendations can be incorrect in specific contexts and cases. However, evidence suggests that XAI systems have not yet had substantial success in facilitating calibrated trust and improving the collaborative Human-AI decision outcomes [2,3,5]. For instance, our recent study revealed a pattern when participants became gradually less interested in the details of the AI explanations and overlooked them during the experiment [5]. Such a failure to aid trust calibration was due to the assumption that XAI users would cognitively engage with explanations and interpret them without bias. Trust calibration as an explanation goal is likely to be achieved when XAI users engage with explanations and analyse its content. This may suggest that presentation and interaction design is as important as the content in XAI to engage users with explanations and persuade them to apply reflective thinking.

We argue that cognitive biases provide a useful lens to understand why the explanations do not eliminate humans' over-trust or under-trust in AI recommendations. For instance, over-trust may be linked to confirmation bias [5]. Confirmation bias represents humans' tendency to seek or hastily believe information that matches their beliefs, values and desires. This bias favours some explanations or parts of them and neglects others [4]. Under-trust, on the other hand, may result from anchoring bias, which occurs when humans look at salient parts, themes or features in an explanation that match a reference point, e.g., similar patterns they encountered in the past, and accordingly judge the quality of AI recommendation to be untrustworthy [5].

In this paper, we state that a successful trust calibration in recommendations supported with XAI interfaces, needs more



(A) Regular XAI interface (B) Friction XAI interface (C) Sample of Local Feature explanation

Fig. 1. Tow conditions. (A) depicts the regular XAI interface. (B) Adds friction element to the regular XAI interface. (C) The explanation presents a list of patient profile features; Green bars represent contributed features to accept the prescription, and red bars shows contributed features that might influence accepting the prescription.

than a correct and impartial explanation content and shall benefit from novel design elements to persuade and nudge users towards more cognitive engagement.

Engaging people in more analytical thinking to reduce the impact of cognitive biases on decision-making has been researched in other domains where their successes can be replicated in the case of XAI for trust calibration. Friction is one promising approach that is based on introducing interactions to hinder people from habitually following certain courses of actions when interacting with technology [8]. Design with friction aims to nudge users by instilling doubts into their automatic behaviour [7]. People tend to become more careful decision-makers when they perceive a certain level of risk as a consequence of their behaviour during the decision-making process [13]. They attempt to break mindless behaviour and prompt analytical thinking. In this paper, we hypothesise that friction induce a higher level of cognitive engagement with AI explanation and, consequently, facilitate the trust calibration process. We present results from a preliminary comparative study with 16 participants, using two XAI designs: a friction-free design, and a design with friction-based digital nudging elements.

II. Research Method

Our aim is to examine whether friction can help trust calibration by nudging humans towards more cognitive engagement during Human-AI collaborative decision-making task. We compare users' interactions with two XAI interface designs. The first design is based on information content only (friction-free), and the second design is augmented with friction (while having an identical interface to the first design). Our research question is:

- Does friction have the potential to increase cognitive engagement with AI explanations during a Human-AI collaborative decision-making task?

A. Case study and interfaces design

We designed mock-up interfaces for a prescription screening AI-based tool. Prescription screening is a process that is utilised in clinics by medical experts to ensure that prescriptions are prescribed for their clinical purpose and fit the patients' profile and history. We chose this case study to reflect an everyday Human-AI collaborative decision-making task where habits formation, biases and, consequently, trust calibration errors are likely to happen. We designed mock-ups based on templates and interfaces familiar to our participants in their everyday decision-making tasks (See Fig.1.). Both interfaces, the friction-free in Figure 1.A and the friction-augmented in 1.B provide explanations to justify the provided AI recommendations. The explanations in Figure 1.C included a list of patients' data features that the AI used for generating the recommendation whether to reject or accept a prescription. Our mock-ups mimicked a web-based tool and were meant to simulate the user experience when working on an actual system. As the participant clicks on a prescription, the tool shows the patient profile and the recommendation (accept a prescription or reject a prescription). The participant can click on a button next to the recommendation to understand the AI rationale of why the prescription should be accepted or rejected. The recommendations were not all similar to what an AI-based algorithm would generate. Instead, we wanted to simulate an uncertain and dynamic nature of AI-based recommendations where trust calibration is a crucial design goal, i.e., we included correct and incorrect recommendations. We designed the recommendations through a collaboration with medical experts, based on the AI system they used. In addition, we designed the recommendations to be non-trivial where explanations are needed to make informed decisions and calibrate users' trust. In total, we have designed ten friction-free interfaces and ten friction-augmented interfaces to be used in this experiment.

B. Conditions

We designed two different sets of user interfaces, friction-free and friction-augmented. Interfaces varied in their approaches of completing the Human-AI task. The conditions were:

- *Friction-free XAI interface.* Participants could complete their Human-AI task by clicking on "*Confirm*" or "*Cancel*" buttons to either accept or reject the AI recommendation.
- *Friction-augmented XAI interface.* (While the interface included identical features to the friction-free XAI interface) The participants were asked to confirm

whether they completed their task with or without viewing the explanation..

In the explanations provided, each patient data feature was paired with a specific feature-based value that estimated the contribution in the AI decision process (see Figure 1.C). The explanation reflected Local Feature Importance explanation as a common approach for explaining AI recommendations [9].

C. Participants

A total of 16 medical practitioners participated in our study after sending invitation emails to three different organisations. Previous knowledge in screening patients' prescriptions was an inclusion criterion. Each participant took an approximation of 15 minutes to complete ten collaborative Human-AI tasks. Our data was collected in two different stages. The first was a part of an earlier study when we asked the participants to complete their Human-AI tasks using the friction-free XAI interfaces set. In the second, we approached the same group of participants after eight months to complete ten Human-AI tasks using the same friction-free XAI interfaces augmented with friction. The time interval helped to eliminate the possible learning effect and also fatigue effects of completing 20 Human-AI tasks in one run, i.e., both the ten friction-free interfaces and the ten with friction. Participants in both stages were asked to make optimal decisions while using our AI-based tool. In addition, participants were told that they could access the explanation of the AI recommendation to understand the AI decision. Each of the 16 participants completed 20 Human-AI tasks (10 for friction-free XAI interface and 10 for XAI interface with friction), which resulted in 320 completed tasks.

D. Measurements

Following Zhang et al. [9] and Wang et al. [10], self-reported trust measurements might not be reliable in Human-AI interaction. Trust calibration shall rather be measured more objectively via behavioural indicators. Hence, we relied on how participants engaged with AI explanations to indicate whether XAI interfaces can help trust calibration. We assumed that participants calibrate their trust better when they interact and read the AI explanations. Automated tracking data were collected each time participants used our mock-ups. The data included the timestamps of when each participant started a Human-AI task and when an explanation was accessed, skipped or exited. We measured participants cognitive engagement with AI explanations using both quantitative and qualitative behavioural indicators:

- *Interaction with Explanations.* This is a binary variable that indicates whether participants accessed the explanation or skipped it.
- *Time spent on Explanations*. This is a numerical variable that measures participants overall time in reading the explanation.
- *Concurrent verbal reports of thinking-aloud.* As participants worked through the tasks, we asked them to think aloud to verbalise their thoughts and reasoning. The goal for such practice was to analyse the participants' perception of friction design elements and its role in their engagement and user experience.

III. Results

A total of 16 participants completed the study in both conditions. Participants performed 320 Human-AI tasks and accessed the explanations 217 times out of 320. In both conditions, participants who accessed explanations spent 12.26 seconds on average viewing them (Std. Deviation 6.07, range 3.0-36.0).

A. Interaction with Explanations

To understand the effect of friction on participants engagement with AI explanations, we analysed *Interaction with Explanation* variable in both conditions. Overall, participants accessed AI explanations were 87 in friction-free XAI condition compared to 130 in friction-augmented XAI design. Distributions of explanation interaction were significantly different across both interface conditions ($\chi 2$=44.835, degrees of freedom 1, corresponding to p=0.0004) using Chi-square test. This means participants in friction XAI interface condition interacted significantly more than participants in friction-free XAI interface design. Furthermore, we observed that as the participants progressed in the study the need to show the cognitive task declined and the participants started to access the explanation without the need for friction. This can be due to learning effect and a way to circumvent the friction as we inducted participants to try to interact in a way similar to the real-world. One another possible interpretation is that internalised that interacting with explanations as an integral part of the Human-AI collaborative decision-making process. This means that XAI interactive design needs to have a guiding role and nudge towards forming a habit. People tend to follow measures and instructions [16] in which the XAI direct users for seeking of explanation. Hence, in addition, to elicit explainability informational content needs for a Human-AI task [12,19], XAI interfaces designers may need to find ways to nudge users to interact with these explanations, till they internalise that behaviour.

B. Time spent on Explanations

Participants spent on average 12.655 sec (Std.=5.97, range=3-33sec) reading the explanation in friction-free interface condition compared to average of 12.007 sec (Std.=6.144, range=3-36 sec) for friction augmented XAI interface. We used repeated One-Way ANOVA test to test whether the time in both conditions was significantly different. Our results showed that participants who used friction XAI interface showed no significant difference in their time spent reading and viewing the explanation compared to friction-free interface F (2,215) =0.592, p=0.443. A possible interpretation is that friction as a nudge-based approach did not encourage participants to spend more time with the explanation and potentially engage cognitively with the explanation. Another interpretation could be that our explanations did not require more time to spend in both conditions. Moreover, nudge being based on reactive thinking (i.e., automatic response to follow certain behaviour) could be another possible explanation to the lack of difference in consumed time. However, for behaviours requiring more cognitive engagement, a follow-up persuasion may need to be implemented.

C. Concurrent verbal reports of thinking-aloud.

We made several observations through analysing think-aloud data aiming to understand how participants engaged with XAI in both conditions.

For friction-free XAI condition, eight participants followed a peripheral route in which they applied heuristics and quick judgments to decide if they would engage cognitively with the explanation. For instance, one participant only checked some parts of the explanation and stated: "*… the average pharmacist does not need to look for all these values*". Friction condition, on the other hand, was more effective to make participants commit to the desired behaviour to follow the instructions and pave the way to apply analytical thinking. One participant mentioned: "*I found the message a bit annoying but after reading the explanation I understand why the AI did that*". Another participant also commented: "*the explanation was quite helpful, and it was interesting that no prior treatment was main factor to the AI … I was about to ignore it*". However, it was also recognised that some participants did not perceive the friction experience positively at all times, rather, they perceived it as an impediment to their task. One participant mentioned: "*it is a bit annoying to have this message every time … this patient case is straightforward*". An implication for friction design may need to be associated with a pattern of undesired behaviour, e.g., a user who skips the explanation all the time or an edge case recommendation where explanation is importantly needed to help calibrate trust.

Finally, although participants in both conditions felt that the explanation was helpful in their Human-AI task, some participants noted that they were sceptical to engage cognitively with the explanation. They mentioned several trust-related issues regarding the explanation and its content. For example, four participants questioned the source of the explanation, and one mentioned, "*I cannot fully trust this explanation, how it is generated and what data sources are used to build this AI*". In general, humans' motivation to cognitively involve with an explanation is affected by explainer competency and experience [11]. As an implication, designing XAI interfaces for trust calibration shall be based on a continuous approach starting from building appropriate trust and increase perception of usefulness and validity of explanations in the task. This is related to earlier work by Cai et al. [18], who developed an onboarding phase to help trust calibration in collaborative Human-AI tools. Their approach was to inform human decision-makers about the capabilities, limitations and data sources of the AI before using the AI-based decision-making tool.

IV. Conclusion and Future Directions

Despite the assertion advantages of explanations in the Human-AI collaborative decision-making tasks, their benefits might be limited due to several design limitations [5], such as skipping or misapplying them. Ultimately, a design that promotes desired behaviour and mitigate errors would fundamentally improve trust calibration. In this paper, we used friction as a nudge-based approach to persuade humans to interact and engage cognitively with AI explanation; thus facilitate a calibrated trust. In this preliminary experiment, our results demonstrated that friction-augmented design had a potential to help engage users more with explanations. However, it did not necessarily lead to a thorough cognitive engagement with their content. Future work would require longitudinal studies and more objective measures to examine whether habitual effect and desensitisation happen in a long-term interaction, e.g., through event-driven diary studies with automatic capture of interaction parameters.

Also, future work concerns ways to make users' engagement with the explanations effective and facilitate correct interpretation. It has been shown that people often over-estimate their understanding of an explanation due to an effect called to the Illusion of Explanatory Depth (IOED) [4]. Further investigation is needed to devise measures to facilitate a calibrated understanding of an explanation. For instance, similar to learning, feedback-based approaches where people self-awareness of their level of knowledge, might help [14]. Feedback can initiate changes to knowledge, behaviour, habits, motivation, and the socio-cultural environment [15]. For instance, the XAI system might present performance metrics such as what a user has learned about the AI through interacting with the XAI interface and what they have potentially missed when they skip it.

Acknowledgment

This work is funded by iQHealthTech and Bournemouth university PGR development fund.

References

[1] Green, B. and Chen, Y., 2019. The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), pp.1-24.

[2] Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F. and Gajos, K.Z., 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Translational psychiatry, 11(1), pp.1-9.

[3] 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.

[4] Oswald, M.E. and Grosjean, S., 2004. Confirmation bias. Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory, 79.

[5] Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N. and, Ali, R., 2021. Explainable recommendations and calibrated trust: two systematic users' errors. The Computer Journal, IEEE (In press). https://staffprofiles.bournemouth.ac.uk/display/journal-article/340390

[6] Kool, W. and Botvinick, M., 2018. Mental labour. Nature human behaviour, 2(12), pp.899-908.

[7] 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).

[8] Mejtoft, T., Hale, S. and Söderström, U., 2019, September. Design Friction. In Proceedings of the 31st European Conference on Cognitive Ergonomics (pp. 41-44).

[9] 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).

[10] Wang, N., Pynadath, D.V. and Hill, S.G., 2016, March. Trust calibration within a human-robot team: Comparing automatically generated explanations. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 109-116). IEEE.

[11] Kolb, D.A., 2014. Experiential learning: Experience as the source of learning and development. FT press.

[12] Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M. and Hussmann, H., 2018, March. Bringing transparency design into practice. In 23rd international conference on intelligent user interfaces (pp. 211-223).

[13] Samuelson, W. and Zeckhauser, R., 1988. Status quo bias in decision making. Journal of risk and uncertainty, 1(1), pp.7-59.

[14] Schunk, D.H., 2003. Self-efficacy for reading and writing: Influence of modeling, goal setting, and self-evaluation. Reading &Writing Quarterly, 19(2), pp.159-172.

[15] Rozenblit, L. and Keil, F., 2002. The misunderstood limits of folk science: An illu-sion of explanatory depth. Cognitive science, 26(5), pp.521-562.

[16] O'Kell, S.P., 1988. A study of the relationships between learning style, readiness for self-directed learning and teaching preference of learner nurses in one health dis-trict. Nurse Education Today, 8(4), pp.197-204.

[17] Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020a, September. Explainable recommendations in intelligent systems: delivery methods, modalities and risks. In International Conference on Research Challenges in Information Science (pp. 212-228). Springer, Cham.

[18] 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.

[19] Naiseh, M., Jiang, N., Ma, J. and Ali, R., 2020b, April. Personalising explainable recommendations: literature and conceptualisation. In World Conference on Information Systems and Technologies (pp. 518-533). Springer, Cham.