Assessing Satisfaction in and Understanding of a Collaborative Explainable AI (CXAI) System through User Studies

Tauseef Ibne Mamun[1], Lamia Alam[1], Robert R. Hoffman[2], Shane T. Mueller[1]
[1]Michigan Technological University [2]Institute for Human and Machine Cognition

Modern artificial intelligence (AI) and machine learning (ML) systems have become more capable and more widely used, but often involve underlying processes their users do not understand and may not trust. Some researchers have addressed this by developing algorithms that help explain the workings of the system using 'Explainable' AI algorithms (XAI), but these have not always been successful in improving their understanding. Alternatively, collaborative user-driven explanations may address the needs of users, augmenting or replacing algorithmic explanations. We evaluate one such approach called "collaborative explainable AI" (CXAI). Across two experiments, we examined CXAI to assess whether users' mental models, performance, and satisfaction improved with access to user-generated explanations. Results showed that collaborative explanations afforded users a better understanding of and satisfaction with the system than users without access to the explanations, suggesting that a CXAI system may provide a useful support that more dominant XAI approaches do not.

INTRODUCTION

The field of Explainable AI (XAI) is an emerging subdomain at the intersection of Human Factors and Artificial Intelligence (AI) with the goal of investigating new ways and methods for explaining complex AI agents to human users. XAI techniques have mainly used algorithmic approaches [e.g., Cotter et al. (2017)] designed to generate information that serves to explain the choices and actions of an AI to its users. These algorithmic approaches often focus on visualization algorithms like LIME (Ribeiro et al., 2016) to explain specific decisions and actions, giving an understanding of how specific features may have led to the outcomes. For example, an image classifier might identify the portions of an image that were most important in leading to the answer, or a medical diagnostic system may visualize which signs and symptoms were most important. However, in all cases, the burden is on the user to incorporate this information into their own understanding, so they must engage in self-explanation to effectively use the XAI output. This suggests that these self-explanations by users---which are already occurring---might be harnessed to provide collaborative explanations to others. This type of explanatory platform can remove the shortcomings that arise from model-intrinsic explanatory algorithms (Das & Rad, 2020) because a change of AI's architecture does not affect the explanatory platform.

Social Q&A (SQA) systems are widely-used software platforms that allow users to work together to resolve issues, find a solution/solutions to a problem, or document and share errors and mistakes of software systems. These include sites like Stack Exchange or Stack Overflow, in which users interact with each other to pose and answer questions, vote on the accuracy of answers, and assign credit for good contributions. Research has shown that both direct and indirect interactions on Social Q&A sites can improve answer quality (Tausczik et al., 2014), suggesting their benefit goes beyond the users creating questions and answers. Thus, researchers advocated that they may be able to be deployed as a human-centered XAI approach (Mamun et al., 2021b). However, because the field of XAI is dominated by algorithmic approaches, the potential practical benefits of SQA systems remain underappreciated.

Mamun et al. (2021b) have proposed the central elements of implementing SQA systems to support AI as *collaborative XAI* (CXAI). A prototype system was developed and used to collect user explanations of an AI image classifier system. The goal was to give users the general advantages of SQA systems while focusing workflow and usability on the particular needs of AI explanations. This system has traditional features of a general social QA platform where users can associate keyword(s) to their posts, bounties to engage users in the platform, also some novel features like a list of topics (Mueller et al., 2019) as 'triggers' that can be used to categorize the postings in the system.

Our goal in this paper is to evaluate the effectiveness of the CXAI system [see Figure1] described in Mamun et al. (2021b). Hoffman et al. (2018a) and Hoffman et al. (2018b) described a comprehensive measurement approach for assessing explanations in the context of AI systems. This included: (1) assessing explanation 'goodness'; (2) measuring user mental models; (3) assessing qualitative measures of trust, satisfaction, and reliance; and (4) measuring human-AI task performance. In terms of explanation 'goodness', Mamun et al. (2021a) showed that the explanations arising from CXAI are written are mostly true, competitive with other human and AI-generated systems in their complexity, and deal primarily with *what*-style explanatory reasoning. In the present study, we will evaluate whether the CXAI system (Mamun et al.,

2021b) improves user mental models/knowledge, satisfaction, and performance insofar as it leads to improved predictive accuracy.

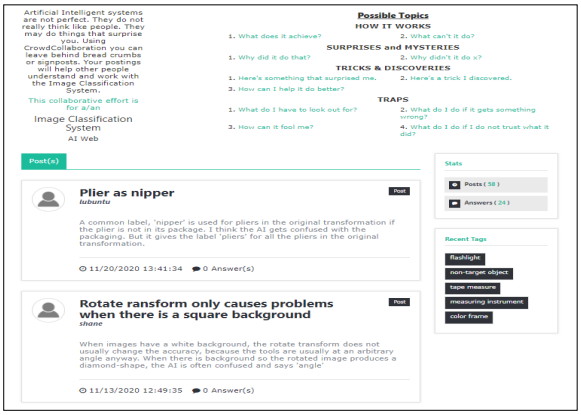

Figure 1. Screenshot of the CXAI system, with two entries related to the tool image classifier. The interface allows users to search by keywords, filter by tags, add new entries, and respond to existing entries with comments or answers.

STUDY OVERVIEW

The CXAI system described by Mamun et al. (2021b) focused on a small image classifier data set, in which an AI system provided labels for images of 10 categories of hand tools under a number of distinct image transforms (e.g., rotation, distortion, black-and-white transforms, etc.). In the two studies presented here, we report tests of comprehension and performance (in Study 1) of users interacting with explanations generated via CXAI, and qualitative assessments of satisfaction by human participant responses to the CXAI system or explanations generated by that tool (in Study 2). In both user studies, we compared the CXAI system (Mamun et al., 2021b) to a Visual Browser (Mueller et al., 2020) of an image classification database [see Figure 2] which enabled users to explore patterns and see the results of the image classifier. This visual browser was the same interface CXAI users had access to when creating CXAI entries. Thus, the control group did not receive explanations per se but were presented with a visual browsing tool that enabled them to make their own discoveries and explanations.

USER STUDY 1: TEST OF COMPREHENSION AND PERFORMANCE

The first study measured whether interacting with explanations in an existing CXAI system would improve user knowledge of the AI system. To do this, we assessed both accuracy and time to complete a set of knowledge questions about particular patterns in the AI system. We hypothesized that if the CXAI system is effective, it should allow users to answer questions about strengths, limitations, and errors in the system better (faster and more accurately) than direct browsing of the image database.

Figure 2. Visual Browser for an image classifier. This allows users to browse different transformations and explore the kinds of errors made for specific images. In the above example, an axe with a leafy frame was classified as 'plant'.

Participants

69 undergraduate students from Michigan Technological University, enrolled in an introductory psychology course participated in the user study in a credit-based compensation structure. Although we did not collect information about their field of study, historically the majority of students enrolled in this class are studying an engineering or science discipline.

Method

A set of questions (10) about the image classifier system performance was created. The questions covered all the transformations of the Visual Browser, and asked participants how the AI would perform for a certain type of tool image in certain conditions. Consistent with other research in this domain (Kulesza et al., 2015) mental model was assessed with knowledge questions specifically designed for this system. Each question required establishing a pattern that related to more than a single example image so that they would require examining multiple images in the browser. The questions were multiple-choice, with three to five options, so that by guessing, accuracy would be expected to be below 50%. The answers to each of the questions could be found in either the Visual Browser or the CXAI system. The experiment was a between-subjects design so that each participant only had access to one of the systems in order to answer the questions. In both conditions, after each question, the participants self-reported whether they actually used the system, or if they guessed to answer the question. After agreeing to the consent form, and answering a few demographic questions, a

participant was trained on a particular system with a video tutorial on the system. After that, the participants answered the questions without time constraints. All procedures were approved via the MTU institutional review board.

Results

Results showed that the users of the CXAI system achieved higher accuracy than the control group (proportion correct of 0.65 and 0.54, respectively; $t(66.67) = -2.21$, $p = 0.03$; $d = 0.56$), which shows that even though the answers were attainable via the browser and the CXAI was developed using the browser, the collaborative entries provided a substantial benefit over the browser—even though it was not available to users of the CXAI system. It is also useful to examine the time needed to answer the questions, as this might make one system more palatable to users than another. Figure 3 shows the distribution of total time taken for each user (in seconds) across participants in each group. A t-test showed no statistically significant difference between total time across conditions: $t(58.6) = -0.93$, $p = 0.24$; $d = 0.23$; and furthermore Kolmogorov-Smirnov test also showed no significant difference between the shape of the distributions: ($D = 0.13$, $p = 0.86$). Thus, these results supported our hypotheses insofar as the users of the CXAI system took a similar amount of time to the users of the Visual Browser to achieve higher accuracy.

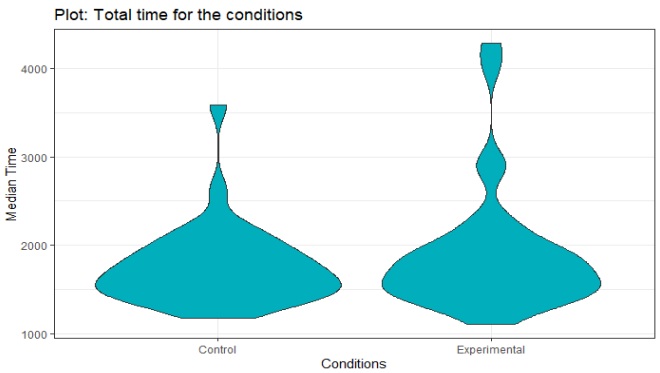

Figure 3. Total time for the conditions

Finally, we examined whether accuracy was impacted by the self-report of whether the participants used the system or guessed. In cases where the user was guessing, no substantial difference existed between the two conditions, and accuracy was around 25%--as expected for the 3-5 item multiple-choice test [see Table 1]. In cases in which the users reported using the tool, the difference in accuracy was even higher (73% vs 55%), which was also statistically significant (t (66.7) = -2.22, $p = 0.003$; $d = 0.54$). However, users were also more likely to report they were guessing in the CXAI condition than in the control (14% vs 5%), which was statistically significantly different according to a Chi-squared test ($X^2(2) = 641.74$, $p < 0.001$.)

This shows users in the experimental condition may have a tendency of trading off accuracy for effort (Liesefeld & Janczyk, 2019), insofar as they were less willing to use the CXAI system even though it improved their chances of answering questions.

Table 1. Mean accuracy for the system use and nonuse

System	System Used	Mean Accuracy
Control (Visual Browser)	Yes (324)	0.55
Control (Visual Browser)	No (16)	0.25
CXAI system	Yes (301)	0.73
CXAI system	No (49)	0.26

Discussion

This user study shows that explanations created via a CXAI system can be used to explain AI systems to users, and it impacts their ability to correctly answer questions about patterns of performance of the AI system. The explanations generated in a collaborative setting are mostly accurate (Mamun et al., 2021a), and users can do statistically better with the CXAI system in contrast to a system that allows them to browse the raw data. The higher accuracy with the CXAI system shows the explanations from the CXAI system enable the participant to form an initial mental model (in this case mostly a correct mental model) of the AI system, and subsequent experience with the system, that can include user-centric expert-generated explanations, would enable a participant to refine their mental model.

Nevertheless, we saw slightly less engagement with the CXAI system, suggesting some users likely perceived that it involved greater effort even if it improved accuracy, (as shown by significant chi-squared test) despite it is not taking significant amount of time (see Figure 3). Thus, it may be that the CXAI system will be viewed by users as poor on Hoffman et al.'s (2018) satisfaction criteria, including subjective assessments of satisfaction, trust, completeness, and sufficiency. In the next study, we examined these more qualitative subjective measures directly.

USER STUDY 2: ASSESSMENT OF SUBJECTIVE MEASURES OF SATISFACTION

Goals

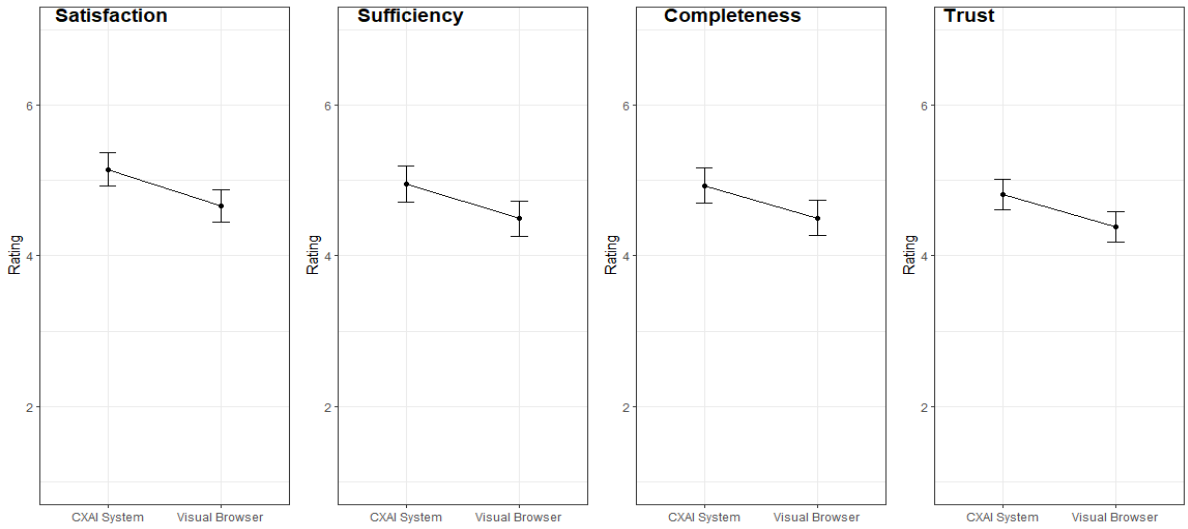

Figure 4. Comparison of the two systems on the attributes (satisfaction, sufficiency, completeness, trust)

Another way of assessing explanations is via subjective measures such as satisfaction, trust, and reliance (Hoffman et al. 2018b). Presumably, in the previous study, users might not notice using the system improved their accuracy because they did not experience the condition without the CXAI tool. Consequently, subjective measures might still be important for predicting the adoption of the tool. Furthermore, Study 1 suggested that users were more willing to guess when using the CXAI system, indicated by the self-report of their using the system. This may be revealed in subjective assessments. Consequently, in this study, we assessed explanations from the collaborative platform using different qualitative measures.

Participants

43 undergraduate students from Michigan Technological University participated in the user study in a credit-based compensation structure.

Method

The participants were given a made-up scenario akin to a use case currently being adopted by retailers; they have been attached to a Hardware Store where two systems are used (Visual Browser and CXAI system) to identify the kind of tool based on a user's image. Each participant used both the CXAI and visual browser and was given 8 questions regarding different instances, transformations, or tools. There were two forms of the study in which the Visual Browser and the CXAI system were counterbalanced across odd and even-numbered questions respectively, to enable a within-subject comparison of the two systems. For each question, a sample of explanations regarding the instance, tool, or transformation was attached from either the CXAI system or Visual Browser. The three best examples determined by the researchers related to a question were given regarding the instance, tool, or transformation for the Visual Browser, and all the explanations that were found during a search in the CXAI system regarding the instance, tool, or transformation were given for the CXAI system for the conditions. Thus, the user did not interact with the interface of either system but was simply presented with best-case information these systems provided. The participants answered the questions with the help of the explanations provided to them for a question. For each question, a participant gave his/her inputs in a 7-point Likert-scale for each attribute (satisfaction, sufficiency, completeness, trust [see (Hoffman, Mueller, et al., 2018b)] where a 7 denotes a positive attitude to an attribute and a 1 denotes a negative attitude to an attribute. All procedures were approved via the MTU institutional review board.

Result

For all the attributes (satisfaction, sufficiency, completeness, trust), CXAI system produced more positive ratings than Visual browser [see Figure 4], and these were all statistically significant: Satisfaction: $t(86) = -4.46$, $p < 0.001$; $d = 0.4$; Sufficiency: $t(86) = -3.88$, $p < 0.001$; $d = 0.36$; Completeness: $t(86) = -3.64$, $p < 0.001$; $d = 0.33$; Trust: $t(86) = -4.17$, $p < 0.001$; $d = 0.32$.

Discussion

This study shows that users found explanations produced by the CXAI system to be generally satisfying, sufficient, complete, and trustworthy, as they all achieved ratings positive ratings of around 5 on a 7-point scale. Furthermore,

these ratings were universally higher than those for example-based explanations based on the browser *which the users of the CXAI system used to generate these explanations.* This suggests such collaborative XAI systems can generate explanations users find helpful.

GENERAL DISCUSSION

The results of the two studies reported here show that collaborative explanations can be helpful, insofar as they help produce accurate answers to questions about the system while not taking substantially longer to answer, and they are also rated as more satisfying, sufficient, complete, and trustworthy in comparison to example-based explanations obtained by browsing the database itself. Notably, the users gather knowledge efficiently from a collaborative environment that is more effective in nature than the system CXAI users used to create the collaborative explanations.

One important caveat is that in the between-participant Study 1, participants self-reported that they guessed about 3 times more often (14% vs. 5%) when using the CXAI system than when browsing the database directly. This may stem from the ease with which some questions could be investigated using the visual database browser, or the challenge of finding relevant CXAI entries related to particular questions. But this is coupled with the relative disadvantage for accuracy—the browser can make it easy to come up with a wrong answer, which might only be detected if a user engages with the browser more intensively in order to explore and establish patterns of behavior. Furthermore, the browser view of the database is not generally available, so the advantage of a CXAI system may be even greater.

Another limitation of this study is that it does not compare the CXAI explanations directly to the kinds of algorithmic explanations often generated by modern XAI systems. The previous examination of the CXAI system (Mueller et al., 2019) concluded that the nature of explanations produced by the system answer very different questions than are typically the target of XAI algorithms. Importantly, CXAI explanations tend to focus on *what*-style questions, whereas algorithmic systems tend to focus on *why* questions: especially focused on local justification of particular decisions. Thus, these different explanatory systems are better thought of as complements to one another, rather than serving as alternative solutions to the same problem.

The CXAI system deliberately resembles SQA systems like StackExchange. Based on our evaluation of this initial prototype, we believe a version of the CXAI system may be best suited for users of a small group of users of an AI system. This might involve an internal team within a company (i.e., as an alternative to a bug-reporting system focused on workarounds and limitations of the tool they use), or a shared community of interest (i.e., radiologists using a particular algorithm for diagnosing particular disorders). In comparison to other SQA systems such as StackExchange, it does not incorporate many of the mechanisms for incentivizing contributions and assessing accuracy or importance of answers, which is critical for those systems because they allow contributions from any interested parties. Furthermore, we believe the strengths of the system come from the targeted use within a context and among a group of workers with a shared mission. Thus, general questions about, for example, convolutional neural networks or PyTorch would probably be better supported by a StackExchange topic which will draw from a broader group of users with more general experience.

ACKNOWLEDGEMENTS

Author contact: tmamun@mtu.edu. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) through the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711. This material is approved for public release. Distribution is unlimited.

REFERENCES

Cotter, K., Cho, J., & Rader, E. (2017). Explaining the news feed algorithm: An analysis of the" News Feed FYI" blog. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1553–1560.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *ArXiv Preprint ArXiv:2006.11371*.

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For "Explainable AI." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 197–201.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *ArXiv Preprint ArXiv:1812.04608*.

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. https://doi.org/10.1145/2678025.2701399

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs (?). *Behavior Research Methods*, *51*(1), 40–60.

Mamun, T. I., Baker, K., Malinowski, H., Hoffman, R. R., & Mueller, S. T. (2021). Assessing Collaborative Explanations of AI using Explanation Goodness Criteria. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 988–993.

Mamun, T. I., Hoffman, R. R., & Mueller, S. T. (2021). Collaborative Explainable AI: A non-algorithmic approach to generating explanations of AI. *International Conference on Human-Computer Interaction*, 144–150.

Mueller, S. T., Agarwal, P., Linja, A., Dave, N., & Alam, L. (2020). The Unreasonable Ineptitude of Deep Image Classification Networks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 410–414.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review,

synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv Preprint ArXiv:1902.01876*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. http://dl.acm.org/citation.cfm?id=2939778

Tausczik, Y. R., Kittur, A., & Kraut, R. E. (2014). Collaborative problem solving: A study of mathoverflow. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 355–367.