Development and Investigation on a Collaborative XAI System (CXAI)

Shane T. Mueller, Ph.D.
Tauseef Ibne Mamun
Michigan Technological University
Robert R. Hoffman, Ph.D.
Institute for Human and Machine Cognition

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

Cite as:

Mueller, S.T., Mamun, T.I., and Hoffman, R.R. (2021). "Development and Investigation on a Collaborative XAI System (CXAI)." Technical Report, DARPA Explainable AI Program.

Abstract

Research using a Naturalistic Decision Making (NDM) approach (Klein, 2008), has suggested parallels between how we explain complex concepts to ourselves and others, and the need for Explainable AI (XAI). From the concept of explanatory reasoning from psychological research and NDM, the concept for Collaborative Explainable AI or CXAI was developed that was reported on a previous DARPA XAI Technical Report, "*Methods for Effective Interaction with XAI Systems: Collaborative XAI (CXAI)*". The current report includes the development of a CXAI system. Also, ways and experiments for assessing an XAI system especially a CXAI system where explanations are generated mostly through self-explanations (Mueller et al., 2021) and collaboration. The goal is to assess this XAI system in terms of explanations, user mental model, and performance.

Outline

1. Pertinent Background

XAI approaches have mainly used algorithmic approaches designed to generate explanations automatically to users. The algorithmic approach mainly focuses on visualization algorithms to explain specific decisions and actions giving an understanding of how specific features may have led to the outcomes. Another way of generating explanations is through human collaboration and self-explanation from which the CXAI concept was created. This type of explanatory platform can remove the shortcoming (Das & Rad, 2020) that arises from model-intrinsic explanatory algorithms because a change of AI's architecture does not affect the explanatory platform. We believe that this collaborative system can enhance and improve existing algorithmic explanation-based systems and provide communities of users with an important resource for understanding a system.

One justification for the usefulness of a collaborative environment for explanation is that it mirrors well-studied frameworks of pedagogy and learning, allowing opportunities for learners to participate irrespective of their experience or knowledge levels. For example, ICAP (Interactive>Constructive>Active>Passive) framework (Chi & Wylie, 2014), suggests the most effective modes for learning involve human-human interactivity where students can better understand a particular topic through dialoguing and explaining to one another. Thus, a collaborative explanation system has the potential to benefit the users at a number of levels, from those who interact with others to create explanations, to those who construct explanations, and those who actively explore the system in order to solve particular problems. Thus, the CXAI system may help users to learn from each other about the AI systems they use. Some of the explanations this can support include: How does an AI system work? What are its shortcomings? What are the reasons for the shortcomings? What are some suggestions, and methods for working around the shortcomings? Thus, the CXAI concept may help provide a user-centric explanation system that does not require algorithms, user models, or complex visualizations, in order to provide important explanations to a user. Furthermore, the explanations elicited may complement those produced by algorithmic approaches, providing a different level of information that is useful and actionable.

1.1 Collaborative Learning

The CXAI system supports human-human learning via collaboration, which has been studied in educational settings. Collaborative Learning has a broad meaning. It can be conducted as a pair or in a group, face-to-face or computer-mediated, synchronous, or asynchronous. However, learning via collaboration can be generally described as a situation in which particular forms of interaction among people are expected to occur, which would trigger learning, although there is no guarantee that the expected interactions will occur (Dillenbourg, 1999).

Learning in collaboration has been suggested to help in developing higher-level thinking skills (Webb, 1982). Students can perform at higher levels when asked to work in collaborative situations than when asked to work individually (Vygotsky, 1980). They also test better when they learn in a collaborative manner (Gokhale, 1995). Students develop valuable problem-solving skills by formulating their ideas, discussing them, receiving immediate feedback, and responding to questions and comments (Johnson, 1971; Peterson & Swing, 1985). Since the XAI approach

advocated here is intended for novel users of an AI system, and one of their goals is gaining problem-solving skills in the context of an AI system, it is promising that collaborative learning has been shown to support these skills in other contexts. Web-based technology is frequently used in the classroom to enrich learning performance, including individual knowledge construction and group knowledge sharing. For example, (Koschmann, 1996) studied web-based collaborative learning systems in the computer-supported collaborative learning (CSCL) paradigm, which are informed by a rich history of cognitive science research about how students learn. Web-based collaborative environments allow equal opportunities for learners to participate without the limitation on knowledge levels (Scardamalia & Bereiter, 1994). Learners in web-based collaborative learning believe it is a time-saving and efficient knowledge-sharing system (Liaw, 2004). In addition, five factors have been found to affect users' attitude towards collaborative web learning (Liaw et al., 2008): system functions, system satisfaction, collaborative activities, learners' characteristics, and system acceptance. The Knowledge Community and Inquiry Model (Slotta & Najafi, 2013) is also relevant, as it involves Web 2.0 technologies where students explore a conceptual domain, express their ideas, and create a collective knowledge base. This type of knowledge base can be used by any future user of an AI system.

1.2 Collaborative Problem Solving

A second collaborative activity supported by the CXAI system is a form of collaborative problem solving: trying to figure out the unknown properties of the AI system together. Problem-solving not only depends on making sense of the behavior of the system but also depends on the division of labor in the group. A number of past systems have been developed to support collaborative problem-solving. For instance, in a web search task, for initial, and synchronous search, a chat-centric view was preferred by 67% of participants in the CoSense tool (Paul & Morris, 2009). This suggests the ability to communicate regarding the problem may be useful for forming explanations about an unknown trait of an AI system as it helps in keeping track of what decisions are made in the group and how each member is performing in the task of problem-solving.

Another important aspect of collaborative problem solving involves how the problem is initially framed and posed to trigger the problem-solving activity. In the initial stage of a collaborative system where problem-solving has not started yet, to initiate problem-solving, specific questions can be useful. Such trigger-questions that initiate explanations include Taxonomic knowledge (What does X mean? What are the types of X?), Sensory knowledge (What does X look like? What does X sound like?), Goal-oriented procedural knowledge (How does a person use/play X?), and Causal knowledge (What causes X? What are the consequences of X? What are the properties of X? How does X affect the sound? How does a person create X?) (Graesser et al., 1996), and similar trigger questions have been examined in the scope of XAI (Mueller et al., 2019).

Collaborative problem-solving tasks also involve both content-free and content-dependent types (Care et al., 2015). Content-free tasks depend on inductive and deductive thinking skills, and content-dependent tasks allow users to draw on knowledge gained through traditional learning areas or subjects. The CXAI mainly supports content-dependent problem solving, because it focuses users on particular cases, errors, and challenges of an AI system. To better enable content-dependent tasks in CXAI, we have implemented specific data fields that allow URL references to

specific problems in the AI system, so that the knowledge can be drawn from these references by the users.

1.3 Motivating Users for Explanations

Our proposed CXAI system is a modified SQA platform (like Stack Exchange or Stack Overflow), but rather than being a general-purpose system for a wide audience, can serve as an explanation system for AI. The goal is to give users the general advantages of SQA systems while focusing workflow and usability on the particular needs of AI explanations. SQA systems often harness the social context in which people ask, answer, and rate content (Oh, 2018), serving as public or community-based resources and relying on natural language communication (Shah et al., 2009) rather than extensive algorithmic data, video, or other means. In order to succeed, however, users of an SQA platform need to be sufficiently motivated to interact with the system. A small community or team may be motivated to communicate intrinsically, but other SQA systems have incorporated specific features that encourage contributions.

For example, some SQA sites vet existing contributions and motivate future contributions by awarding points to users (Oh, 2018). SQA sites typically do not enlist professional or expert answers, though several SQA sites have allowed users to build a reputation within a particular question category and become known as an expert on the site (Shah et al., 2009). A user contributes his/her knowledge because of factors including the user's reputation, self-presentation, peer recognition, etc. (Jin et al., 2015). Motivating users to contribute is important because, along with having more information, the best answers in an SQA platform are correlated with the consistent participation of users, which can be motivated through points (Nam et al., 2009) or bounties (Zhou et al., 2020).

1.4 XAI Evaluation Process

Any new type of explanatory platform must be properly evaluated to ensure its success when deployed. Hoffman et al. (2018a) and Hoffman et al. (2018b) (see also Mueller et al. (2021)) 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. Many systems have been developed with these three criteria in mind, it is actually rare for a system to be evaluated according to them. Although user testing to evaluate human performance remains a gold standard evaluation for AI systems, other measures are equally important for the evaluation of an explanatory system. Besides performance, the comprehensive measurement approach will evaluate the quality of explanations and the user’s mental model. The comprehensive measurement will ensure that an XAI system will work in practice.

2. HUMAN-CENTRIC SYSTEM DESIGN AND DEVELOPMENT

To identify the critical elements of a web-based novel explanatory system similar to a social QA platform, we engaged in a collaborative design effort in which members of our research group worked with an initial system to pose and answer explanatory questions about an AI system, and iteratively refined the interface based on this activity. The system has traditional features of a

general social QA platform (like StackOverflow or StackExchange) where users can associate keyword(s) to their posts, and also some novel features like a list of topics that can be used to categorize the postings in the system. These topics would be the "triggers" (see Figure 2.1) for explanations that have been revealed in the research on the importance of users' goals and needs regarding explanations (Mueller et al., 2019). These topics can also be used in initiating problem-solving discussed earlier. Once one or more topics were selected, it would serve as metadata to contextualize the user's notes and the responses from other users. This would support other users' subsequent searches through the collaborative system. Thus, the artifacts of system development that we examine are not part of a comprehensive user test from a naive user group but may still be informative.

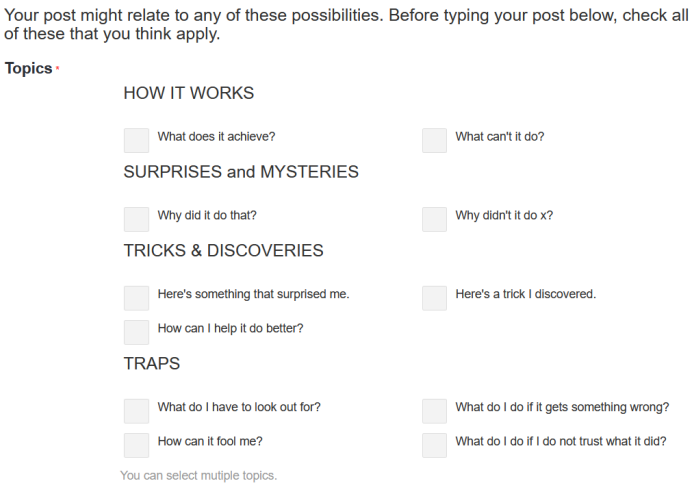

Figure 2.1. Topics as 'triggers.'

Another feature we incorporated through the team's feedback is the ability to add URL reference(s) to their posts about the AI system so that other users can understand the posts with the help of the reference link(s). Another is the use of keywords: if a user wanted to create a new post, this could be associated with one or more keywords and topics that help in categorizing and searching posts. The system has gone through a usability evaluation (Experience, n.d.) after development. Think aloud protocol (Jääskeläinen, 2010) was also used to identify usability issues for the system.

3. OVERVIEW OF CXAI EVALUATION PROCESS

The comprehensive measurement (see section *XAI Evaluation Process*) approach can be done in two steps; an evaluation of the goodness criteria might provide an early heuristic formative evaluation that can be useful for refining the design of the system, without requiring complex or costly human user evaluation of an incomplete system. The rest of the measurements are done after completion of the system with human participants evaluating the system. Next section will discuss the formative evaluation part, after that the following sections will evaluate the system through user studies.

4. THE GOODNESS CRITERIA FOR HEURISTIC EVALUATION

Many designs and evaluation criteria have been proposed by XAI researchers. Among these, Hoffman et al. (2018b) and Hoffman et al. (2018a) identified three kinds of measures obtainable by assessing users of a system: *satisfaction* measures encompassing subjective feelings of trust and reliance and the like; *mental model* metrics related to knowledge and understanding; and *performance*, related to how the human-AI system accomplishes a joint task. They also identified a fourth class of measures described as 'explanation goodness'--criteria often posed by XAI designers as a priori properties of good explanations.

For example, some have argued that an explanation must be **accurate** or **correct**; otherwise, it will hurt users' trust in the system (Papenmeier et al., 2019). Another such property is **scope or focus** (see Doshi-Velez & Kim, 2017; Wick & Slagle, 1989), describing whether an explanation refers to specific cases (local) or large-scale patterns and operations of the system (global). Alam (2020) showed how this scope can impact different aspects of satisfaction, and so it is important for heuristic evaluation. Related to this is **explanation form**, determined by the kind of question the explanation answers. Many XAI systems use justification, which answers a *why* question about the system, justifying why a decision was made or not made. Others have described the goal of **simplicity** (e.g., Kulesza et al., 2015). This can be assessed in several ways, and we will use measures of readability to provide insight into this criterion. Finally, we will examine the extent to which explanations provide workable **knowledge** to the users, rather than just opinions about an AI system. Evaluate the CXAI system with this set of new criteria to determine the strengths of the system as an XAI system, and to provide an example evaluation approach for examining future XAI systems.

The CXAI system itself was developed using Laravel. To populate the system, we created an online browser that allowed users to explore how a popular commercial image classifier performed on a set of 50 images of ten hand tools under several image transforms (see Mueller et al., 2020, which examined the performance of the system). The overall system was developed collaboratively with a set of users including the design team and interested graduate students enrolled in a human factor graduate program as part of their coursework, who were asked to explore the AI system and use the CXAI system to identify errors, patterns, and other issues with the system. Once complete, a few observations were removed from the set as they were mainly 'curiosity' observations and could not be treated as explanations for the AI system by the researchers, resulting in a final set of 43 text-based explanations that we examined further. To these, we added 15 explanations generated by the users of another image classification system that were judged to not be true of the target system. Two independent coders converted this set into 113 independent codable chunks (95 target and 18 foils) based on a set of criteria relating to whether the explanation included multiple independent statements. This unitization process involved one rater who divided statements into independent clauses and a second rater who approved the division. When there were disagreements, the raters discussed and came to a consensus about the unitization. The raters were naïve to the purpose and goals of the paper removing any bias from their work.

4.1. The Knowledge Base Criterion

One goal of explanation for an AI system is to provide a good knowledge base to allow users to engage in self-explanation, sensemaking, and discovery. One concern of the CXAI system is that entries will not be factual, but opinions or other non-factual perspectives, which would reduce the usefulness of the explanations. Consequently, this criterion assesses the extent to which explanatory statements provide that knowledge or might be considered opinion.

The coding of knowledge was done concurrently with the coding of accuracy (the next criterion). Two coders independently coded the 95 original chunks based on whether each statement was an opinion or a factual statement. In total, 77 of 95 chunks were selected based on a set of inclusion criteria. These 77 statements were coded by two independent raters as factual (whether correct or incorrect) or opinion. The raters achieved a moderate level of agreement with κ=.67 (McHugh, 2012). Out of 77 statements, raters agreed on 61 statements as factual knowledge and 9 statements as opinion. For the remaining 7 statements, raters were not in agreement.

This analysis reveals that most statements in the CXAI system relate to factual elements of the AI system, and thus form a reasonable knowledge base for understanding the system. Algorithmic XAI systems are unlikely to produce explanations that appear to be opinions, but they may produce artifacts that users do not consider knowledge-building, and similar coding may help understand the proportion of explanations in an algorithmic XAI system that provide useful knowledge. Importantly, the opinion statements tended to be 'should' statements—advice about how the AI should be used or improved, which may be useful even if it is not factual.

4.2. The Accuracy Criterion

One might expect that novice users will provide explanatory statements that are often incorrect. Consequently, we coded the accuracy of explanations with two independent raters who examined each statement, evaluated it against the results of the actual AI system, and judged its correctness.

To measure accuracy, two independent raters examined each statement and coded it as correct, incorrect, or partially correct, providing justifications when necessary. This included 97 (79 original and 18 foils) chunks out of 113 chunks based on a set of inclusion criteria to establish how many of the statements are codable for accuracy (e.g., removing opinion).
The raters achieved a moderate level of agreement on the cases (weighted κ=0.76). Of the 79 target statements (see Table 4.1), the coding resulted in a total of 66 statements judged correct by both raters, 1 as incorrect by both raters, and 12 in which at least one rater judged it as partially correct (3 of these cases the other rater also judged it partially correct). A Chi-squared test of independence showed that the correctness coding depended significantly on the target/foil distinction ($X^2(2) = 58, p < 0.001$), which demonstrates that the raters were able to discriminate accuracy, and thus that the target explanations achieved a high level of accuracy.

Table 4.1. Number of statements about the AI system (target) vs. Those about another system (foil) coded as correct, incorrect, or with at least one rater judging it partially correct.

	Correct	Partial Correct	Incorrect
Target statements	66	12	1
Foil statements	1	6	11

Consequently, this demonstrates that surprisingly, a group of users can work together, through a collaborative tool, to share accurate explanations about an AI system they are mostly unfamiliar with. Thus, it provides a factual knowledge base that allows users to understand how the system performs.

4.3. The Scope Criterion

Several measures contribute to assessing the scope of explanations. In general, we refer to scope as the extent to which an explanation provides a global description of the system versus an account of a single action. To measure scope, coders examined each statement, and determined, roughly, how many instances in the data set the explanation referred to. Each statement was coded as either referring to a single image in a transformation, 2-5 images in a transformation, multiple images of multiple tools in a transformation (up to 50 images), or multiple transformations in the image classifier (entailing more than 50 images). Two coders independently rated the 79 cases described earlier, producing a moderate level of agreement on these cases, (κ=0.57). The result is summarized in Table 4.2.

Table 4.2. Agreement measures on the coding of explanatory scope. Results suggest most explanations refer to global patterns across multiple image instances, transforms, and categories.

Codes	Both Agreed	Not Agreed
A single image in a transformation	1	2
2-5 of the same images in a transformation	10	2
Multiple images of multiple tools in a transformation	36	7
Multiple transformations	12	9

Though the coders did not achieve strong agreement between them, out of 79 statements, almost all statements were deemed to refer to more than a single case. 64 statements were deemed to refer to multiple images of multiple tools in a transformation, or multiple transformations to connect a statement with their findings. The majority of explanations referred to patterns across multiple

images and tool categories. Thus, explanations in the CXAI tend to be at a much broader scope than most algorithmic XAI systems achieve, insofar as they focus on single cases one at a time.

4.4. The Explanation Form Criterion

Researchers in XAI have often described taxonomies of explanation form (see, Swartout & Moore, 1993). One popular taxonomy was described by Lim et al. (2009), which identifies five basic questions explanations answer: What, why, why not, what if, and how to. To evaluate explanation type, two independent coders coded 95 original chunks to see if each chunk answered one of these questions. If a chunk did not answer a question, the case was rated as 'none'.

Results indicated that independent raters achieved a moderate level of agreement on the cases, unweighted κ=0.76. The result from their coding can be summarized in Table 4.3, which demonstrates that the CXAI explanations mostly answered 'what' questions.

Table 4.3. Coding Result – Intelligible Questions

	What	Why	How To	None
What	69	1	1	3
Why	2	9	0	1
What If	0	0	1	0
How To	0	0	2	0
None	0	0	0	6

These codes are related to the so-called explanation triggers identified by Mueller et al. (2019) (see Figure 4.1). The design of the CXAI system encouraged users to select one or more of these reasons when a new explanation is entered. We compared the form codes to the user-specified trigger codes (see Table 4.4). Results show that the reasons people gave for different explanations varied widely, and although the majority of explanations fall into a 'what'-style explanation type according to Lim et al. (2009), these 'what' explanations appear to have many different purposes, especially describing surprising results, warning others about mistakes, and advising how to handle certain cases. Notably, relatively few statements answer 'why' or 'why-not' questions—and these represent justification-style explanations that are probably the most typical explanations that exist in current XAI systems. However, there were substantial numbers of explanations identified by the users as answering 'why questions' that were coded as 'what' explanations. This may be because the explanations were cued by asking the 'why' question but did not provide a 'why' answer.

Table 4.4. Comparing Triggers with Intelligible Questions. Each rater's coding along the explanation type is shown so that each chunk accounts for two entries in the table.

Triggers	What	Why	Why not	What if	How to	None
Here's a trick I discovered.	10	0	0	0	0	0

Here's something that surprised me.	33	6	0	0	0	3
How can I help it do better?	7	3	0	0	2	2
How can it fool me?	6	0	0	0	0	0
What can't it do?	39	4	0	1	6	4
What do I do if it gets something wrong?	6	0	0	1	1	0
What do I have to look out for?	14	0	0	1	5	0
What does it achieve?	16	3	0	0	1	2
Why did it do that?	39	9	0	0	1	7
Why didn't it do x?	52	4	0	0	1	7

4.5. The Simplicity Criterion

The simplicity of an explanation can be evaluated in a number of ways. For example, explanatory statements could be coded for the number of elements or relations they use. This would be partially related to the scope criterion examined earlier. It could also be coded with detailed mapping of an argument structure, which could also be informative. For the present analysis, we chose to examine some simple textual measures of readability. This criterion will help understand whether explanations made by users for other users—without explicit instruction to create simple explanations---are likely to be comprehensible and understandable.

To measure readability, we used the Flesch reading ease (Flesch, 1946) and Flesch–Kincaid grade level (Kincaid et al., 1975) measures, implemented in readability function in the library 'sylcount' library (Schmidt, 2020) of the R statistical computing platform.

One explanatory statement was removed out of 43 statements because the analysis function failed on the statement. For the remaining observations, the mean Flesch–Kincaid grade level score was 6.48, meaning a reader needs a grade 6 level of reading or above to understand the statements. An alternate score, the Flesch reading ease score produced a mean value of 69.6 (with higher values meaning greater ease). Both of these measures had broad distributions indicating a substantial variation in readability, but they both showed that the statements of the XAI system have an acceptable reading level and most US adults can read them (Huang et al., 2015).

Comparing with other explanatory texts. To compare the simplicity of the CXAI explanations with other explanations, we examined a corpus of explanations collected from the internet, popular press, and other sources (Klein et al., 2019) about general topics. These explanations covered many kinds of complex systems outside of the AI domain. These statements produced a mean Flesch–Kincaid grade level score was 5.17, and a mean Flesch reading ease score of 74.6. Two independent-samples t-tests showed that these explanations were marginally simpler than those produced by the CXAI system (grade level: $t(60.4) = 2.73$, $p = 0.008$; reading ease: $t(52.9) = -2.19$, $p = 0.03$ respectively).

As a second comparison, we selected 10 posts on deep learning from Stack Exchange (*Hot Questions - Stack Exchange*, n.d.). The mean Flesch–Kincaid grade level score for this text was

8.64 and the mean Flesch reading ease score was 53.74, which were significantly less readable than the CXAI explanations (for grade level: $t(12.5) = -2.18$, $p = 0.049$; for reading ease: $t(11.34) = 2.63$, $p = 0.023$).

Finally, we conducted the same analysis on explanations reported in Figure 5 of Hendricks et al. (2016). For these statements, the mean Flesch–Kincaid grade level score was 6.9, and the mean Flesch reading ease score was 81.8. Two independent-samples t-tests showed that these explanations were marginally simpler than those produced by the CXAI system (grade level: $t(53.4) = -0.86$, $p = 0.39$; reading ease: $t(55.6) = -6.5$, $p < 0.001$).

Together, this suggests that the explanations produced via CXAI are written simply at a highly readable level (see Figure 4.1). The readability is simpler than similar explanations of deep learning algorithms, but not quite as simple as explanations produced for in the popular press and online message boards, slightly more complex than AI-generated text explanations.

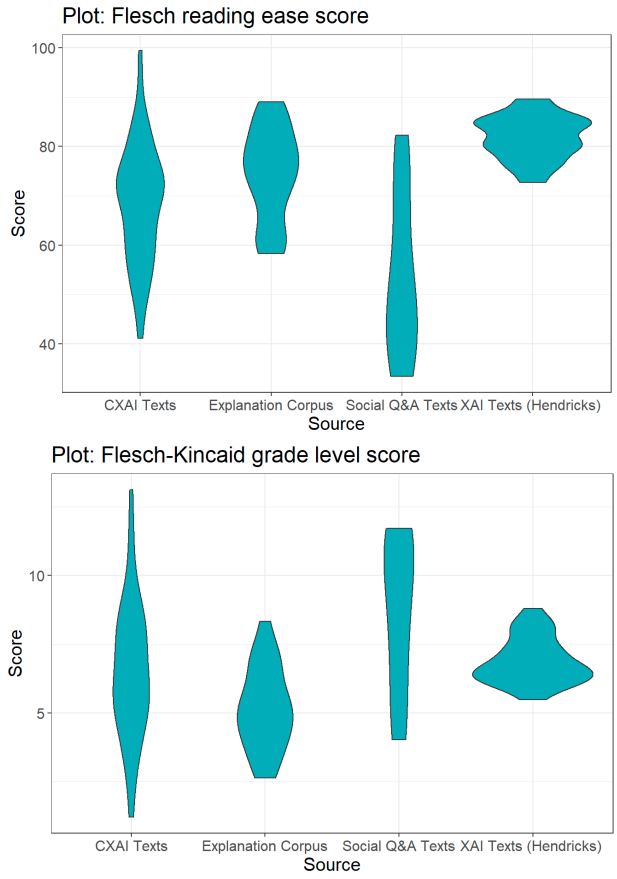

Figure 4.1. Distributions of Flesch Reading ease scores (top panel) and Flesch-Kincaid Grade Level Scores (bottom panel) for CXAI explanations in comparison to three other explanation corpora.

4.6. Discussion of Evaluation by Goodness Criteria

This analysis demonstrates how a heuristic evaluation can be made for an XAI system using the so-called "goodness" criteria, providing a formative evaluation of the strengths and weaknesses of the system. This was achieved with objective measures (such as readability) along with human coding of an explanation case base against criteria such as correctness and scope.

The results of the evaluation showed that the human-generated explanations created in the CXAI system were mostly accurate, knowledge-centric that covered a large scope of an AI system, despite them being generated by relative novices. Furthermore, they were written at an understandable level comparable to other human-generated explanations of general topics and as good or better than human explanations of AI systems and AI-generated explanations.

One important result from this analysis is examining the explanation type. When examined against Lim et al.'s (2009) five categories, most explanations were categorized as 'what' explanations. When we examined the scope of explanations, these what-style explanations tended to describe large-scale patterns across multiple images. This contrasts with most XAI algorithms that are designed to answer local why-style questions in the form of justifications for a specific decision. As such, the CXAI may not provide the same information as those systems but may enable other understanding. Finally, when we examined which explanation triggers these explanations correspond to, we see a great variety of goals and purposes within the 'what' category.

We suggest that these criteria can also be used in other XAI systems as well---especially those that generate algorithmic explanations. However, many such systems will need to develop specific ways of evaluating these criteria based on their explanation types. Furthermore, this analysis points out that the strengths of the CXAI system may differ from other XAI systems, insofar as it tends to capture larger patterns, they tend to be descriptive rather than causal, and they are triggered by a wide variety of reasons.

5. EVALUATION THROUGH USER STUDIES

5.1 Overview

Explainable AI or XAI represents an important category of Human-AI interaction that attempts to improve human understanding and trust in machine intelligence and automation by providing users necessary information that explains algorithms, decisions, actions, and plans. Solutions have been mostly dependent on algorithmic approaches for explaining artificial agents to humans, although some researchers (e.g., Mueller et al., 2019) proposed non-algorithmic approaches via collaboration for explaining AI systems. In this paper, we evaluate one such approach to examine its likelihood of success, using human studies to assess performance, quality of explanation, and user's mental model during use of the system. Results suggest that a collaborative explanation system is helpful, and likely to provide information and support that more dominant XAI approaches do not.

5.2 Background

The field of Explainable AI (XAI) is an emerging subdomain in the domain of Artificial Intelligence (AI) that is investigating new ways and methods for explaining complex AI agents to

human users. XAI approaches have mainly used algorithmic approaches designed to generate explanations automatically to users. These algorithmic approaches mainly focus on visualization algorithms 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 that they must engage in self-explanation to effectively use the XAI output. This suggests that these self-explanations---which are already occurring---might be harnessed to provide collaborative explanations to others. This type of explanatory platform can remove the shortcoming (Das & Rad, 2020) that arises from model-intrinsic explanatory algorithms because a change of AI's architecture does not affect the explanatory platform.

To explore the potential of collaborative XAI, we have developed a prototype system and used this to collect user explanations of an AI image classifier system. Here we describe steps we have taken to evaluate the effectiveness of a collaborative XAI system. (see Hoffman et al., 2018ab; Mueller et al., 2021) 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. Many systems have been developed with these criteria in mind, it is actually rare for a system to be evaluated according to them. Although user testing to evaluate human performance remains a gold standard evaluation for AI systems, other measures are equally important for the evaluation of an explanatory system. Besides performance, the comprehensive measurement approach will evaluate the quality of explanations and the user's understanding and ability to predict performance of the AI system (mental model). AIM and systems.

The explanatory tool focused on a small image classifier data set, in which an AI system provided labels for images of 5 examples of 10 hand tools under a number of distinct image transforms (e.g., rotation, distortion, black-and-white transforms, etc.) Previously, an evaluation of the goodness criteria was done (see previous section of this report), but this assesses explanations independent of users.. In the two studies presented here, we report (in Study 1) qualitative assessments of satisfaction by human participant responses to the CXAI tool or explanations generated by that tool,, and (in Study 2) tests of comprehension and performance. In both user studies, we compared the CXAI system to a visual browser of an image classification database, which enabled users to explore patterns and see results of the image classifier[6]. Both of these systems are depicted in Figure 1 Thus, our 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.

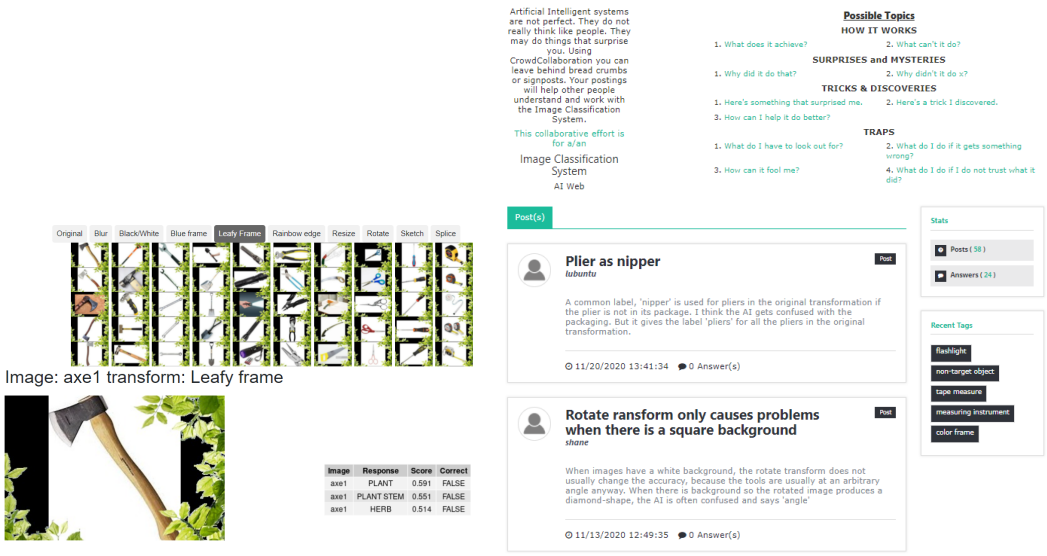

Figure 5.1. Depiction of the AI browser (left) serving as the control condition, and the CXAI (right) serving as the experimental condition.

5.3. User study 1: Test of comprehension and performance

Goals

The first study measured whether the CXAI system would improve user knowledge of the AI system. To do this, we assessed 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.

Participants

69 undergraduate students from MTU participated in the user study in a credit-based compensation structure.

Method

In the user study 1, a set of questions (10) about the image classifier system performance was created. The questions covered all the transformations of the visual browser. The questions asked the participants how the AI would perform for a certain type of tool in certain conditions. Each question has more than one picture of tools that were related to the question. The questions were multiple choice, with three to five answers, 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 AI Database Browser or the CXAI Tool. The experiment was a between-subjects design, so that each participant only had access to one of the system in order to answer the questions. In both conditions, after each question, the participants self-reported whether they used a particular system or 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.) It is also useful to examine the time needed to answer the questions. Figure 2 shows the distribution of total time 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 total distributions: (D = 0.13, p = 0.86). Thus, these results supported our hypotheses insofar as the users of the CXAI Tool took a similar amount of time to the users of the AI Database Browser to achieve higher accuracy.

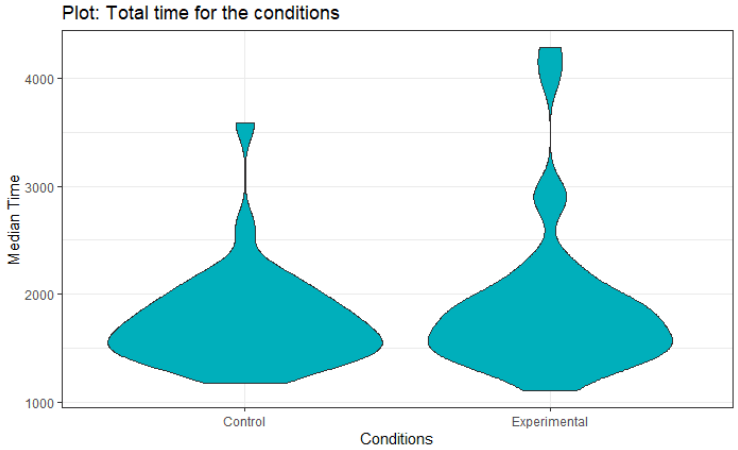

Figure 5.2. Total time for the conditions

Finally, we examined how 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). 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 showed a tendency of trading off accuracy for effort (Liesefeld & Janczyk, 2019) as AI Database Browser is easy to browse. Despite this, if we examine only the 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).

Table 5.1. Mean accuracy for the system use and nonuse.

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

This user study shows that the CXAI Tool can be used to understand AI systems even other explainable AI systems, if necessary. The explanations generated in a collaborative setting are mostly accurate [10], and users can do statistically better with the CXAI Tool in contrast to a system that has visual examples.

5.4. User study 2: Assessment of Qualitative Measures

Goal

Another way of assessing explanations is via subjective measures such as satisfaction, trust, and reliance [4]. Presumably, users might not notice improvements in accuracy, and so subjective measures might be important for predicting adoption of the tool. Furthermore, Study 1 suggested that users were more willing to guess when using the CXAI system, presumably because the perceived effort involved was burdensome. 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 MTU participated in the user study in a credit-based compensation structure.

Method

In user study 2, the participants were given a made-up scenario where a participant has been attached to a Hardware Store where two systems are used (AI Database Browser and CXAI Tool) to explain Hardware Store AI's decision to customers. Unlike Study 1, the experimental design was within-participant, so that each participant used both the CXAI and control tools. The participants were given 8 questions regarding different instances, transformations, or tools (see [6]). There were two counterbalancing conditions (condition 1 and condition 2). In 'condition 1', a participant answered odd number questions using AI Database Browser, and even number questions were answered using CXAI Tool and this was vice-versa for a participant in 'condition 2'. For each question, a sample of explanations regarding the instance, tool, or transformation was attached from the CXAI Tool or AI Database Browser. The three best examples determined by the researchers related to a question were given regarding the instance, tool, or transformation for the AI Database Browser, and all the explanations that were found during a search in the CXAI Tool regarding the instance, tool, or transformation were given for the CXAI Tool for the conditions. 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 [4], where a 7denotes a positive attitude to an attribute and a 1 denotes a negative attitude to an attribute, and a 4 denotes neutrality to the attribute for the question.

Results

For all the attributes (satisfaction, sufficiency, completeness, trust), CXAI Tool produced more positive ratings than AI Database Browser (see Figure 3), 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: : $t(86) = -4.17$, $p < 0.001$; $d = 0.32$.

5.5. Discussion of Human Evaluation

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 a system with visual examples which is the backbone of many XAI systems. One important caveat is that in the between-participant study 1, participants self-reported that they guessed about 3 times more often (15%) 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. In general, the browser view of a database is not available, and questions would have to be answered in very different ways.

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. Our previous examination of the CXAI system [10] 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.

6. REFERENCES

Care, E., Griffin, P., Scoular, C., Awwal, N., & Zoanetti, N. (2015). Collaborative problem solving tasks. In *Assessment and teaching of 21st century skills* (pp. 85–104). Springer.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

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

Dillenbourg, P. (1999). *What do you mean by collaborative learning?*

Experience, W. L. in R.-B. U. (n.d.). *10 Usability Heuristics for User Interface Design*. Nielsen Norman Group. Retrieved May 19, 2021, from https://www.nngroup.com/articles/ten-usability-heuristics/

Gatian, A. W. (1994). Is user satisfaction a valid measure of system effectiveness? Information & management 26, 3 (1994), 119–131.

Gokhale, A. A. (1995). Collaborative Learning Enhances Critical Thinking. *Journal of Technology Education*, *7*(1).

Graesser, A. C., Baggett, W., & Williams, K. (1996). Question-driven explanatory reasoning. *Applied Cognitive Psychology*, *10*(7), 17–31.

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*.

Jääskeläinen, R. (2010). Think-aloud protocol. *Handbook of Translation Studies*, *1*, 371–374.

Jin, J., Li, Y., Zhong, X., & Zhai, L. (2015). Why users contribute knowledge to online communities: An empirical study of an online social Q&A community. *Information & Management*, *52*(7), 840–849.

Johnson, D. W. (1971). Effectiveness of role reversal: Actor or listener. *Psychological Reports*, *28*(1), 275–282.

Klein, G. A. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460.

Koschmann, T. D. (1996). *CSCL, theory and practice of an emerging paradigm*. Routledge.

Liaw, S.-S. (2004). Considerations for developing constructivist web-based learning. *International Journal of Instructional Media*, *31*, 309–319

Liaw, S.-S., Chen, G.-D., & Huang, H.-M. (2008). Users' attitudes toward Web-based collaborative learning systems for knowledge management. *Computers & Education*, *50*(3), 950–961. https://doi.org/10.1016/j.compedu.2006.09.007

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., Hoffman, R. R., & Mueller, S. T. (2021). Collaborative Explainable AI: A non-algorithmic approach to generating explanations of AI. In *Proceedings of the International Conference on Human-Computer Interaction*. New York: Association for Computing Machinery.

Mamun, T., I, Baker, K., Malinowski, H, Hoffman, R. R., & Mueller, S. T. (Forthcoming 2021.) In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

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*.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of Explanation in Human-AI Systems. *ArXiv Preprint ArXiv:2102.04972*.

Nam, K. K., Ackerman, M. S., & Adamic, L. A. (2009). Questions in, knowledge in? A study of Naver's question answering community. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 779–788.

Oh, S. (2018). Social Q&A. In P. Brusilovsky & D. He (Eds.), *Social Information Access: Systems and Technologies* (pp. 75–107). Springer International Publishing. https://doi.org/10.1007/978-3-319-90092-6_3

Paul, S. A., & Morris, M. R. (2009). CoSense: Enhancing sensemaking for collaborative web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1771–1780.

Peterson, P. L., & Swing, S. R. (1985). Students' cognitions as mediators of the effectiveness of small-group learning. *Journal of Educational Psychology*, *77*(3), 299.

Scardamalia, M., & Bereiter, C. (1994). Computer Support for Knowledge-Building Communities. *The Journal of the Learning Sciences*, *3*(3), 265–283.

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social Q&A. *Library & Information Science Research*, *31*(4), 205–209.

Slotta, J. D., & Najafi, H. (2013). Supporting collaborative knowledge construction with Web 2.0 technologies. In *Emerging technologies for the classroom* (pp. 93–112). Springer.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard, MA: Harvard University Press.

Webb, N. M. (1982). Group composition, group interaction, and achievement in cooperative small groups. *Journal of Educational Psychology*, *74*(4), 475.

Zhou, J., Wang, S., Bezemer, C.-P., & Hassan, A. E. (2020). Bounties on technical Q&A sites: A case study of Stack Overflow bounties. *Empirical Software Engineering*, *25*(1), 139–177.