2021

INVESTIGATING THE IMPACT OF ONLINE HUMAN COLLABORATION IN EXPLANATION OF AI SYSTEMS

Tauseef Ibne Mamun
Michigan Technological University, tmamun@mtu.edu

INVESTIGATING THE IMPACT OF ONLINE HUMAN COLLABORATION IN EXPLANATION OF AI SYSTEMS

By

Tauseef Ibne Mamun

A THESIS

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2021

This thesis has been approved in fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Thesis Advisor:	*Shane T. Mueller*
Committee Member:	*Robert R. Hoffman*
Committee Member:	*Jon Sticklen*
Department Chair:	*Kelly S. Steelman*

Table of Contents

List of Figures

List of Tables

Author Contribution Statement

In this thesis, a version of Chapters 2 and 3 has been published as a conference paper (Mamun, Hoffman, et al., 2021). A version of Chapter 4 also has been published as a conference paper (Mamun, Baker, et al., 2021). Chapters 5 and 6 are being prepared for publication.

Abstract

An important subdomain in research on Human-Artificial Intelligence interaction is Explainable AI (XAI). XAI aims to improve human understanding and trust in machine intelligence and automation by providing users with visualizations and other information explaining the AI's decisions, actions, or plans and thereby to establish justified trust and reliance. XAI systems have primarily used algorithmic approaches designed to generate explanations automatically that help understanding underlying information about decisions and establish justified trust and reliance, but an alternate that may augment these systems is to take advantage of the fact that user understanding of AI systems often develops through self-explanation (Mueller et al., 2021). Users attempt to piece together different sources of information and develop a clearer understanding, but these self-explanations are often lost if not shared with others. This thesis research demonstrated how this 'Self-Explanation' could be shared collaboratively via a system that is called collaborative XAI (CXAI). It is akin to a Social Q&A platform (Oh, 2018) such as StackExchange. A web-based system was built and evaluated formatively and via user studies. Formative evaluation will show how explanations in an XAI system, especially collaborative explanations, can be assessed based on 'goodness criteria' (Mueller et al., 2019). This thesis also investigated how the users performed with the explanations from this type of XAI system. Lastly, the research investigated whether the users of CXAI system are satisfied with the human-generated explanations generated in the system and check if the users can trust this type of explanation.

1 Introduction

1.1 Problem Statement

Recent advances in AI have created technology that is both more capable and particularly difficult to understand or predict than previous forms of AI. As applications of AI expand, it has become critical to develop explanatory systems (i.e., Explainable AI - XAI) that will help users to understand and work with these new AI systems. XAI is situated between machine and human and helps humans to understand the machine.

XAI systems are mainly algorithm-focused and often implement untested ideas about explainability, notions that are not informed by the literature of cognitive and educational psychology (Mueller et al., 2021). An alternate approach is needed for generating explanations without depending on algorithms. Research using a Naturalistic Decision Making (NDM) approach (G. A. Klein, 2008) has suggested parallels between how we explain complex concepts to ourselves and others and the needs for XAI. This work suggests a potential role for collaborative explaining and the use of collaboration during the exploratory process. This paved the way for collaborative XAI (CXAI), in which users pose questions and generate their own explanations through collaboration that will help the group to understand the AI system. The hypothesis is that this collaborative system can enhance and improve existing algorithmic explanation-based systems and provide communities of users with an important resource for understanding an AI system. There is a possibility that if we can leverage the knowledge of a team to help them share the discoveries they made, it can help both regular AI users who don’t have access to

explainability and help other explainable system users to better communicate their information throughout a team.

One justification for the usefulness of a collaborative environment for self-explanations is that it mirrors well-studied pedagogy and learning frameworks, allowing learners to participate irrespective of their experience or knowledge levels. Chi & VanLehn (1991) found that learners gain both inductive and deductive knowledge by self-explaining. Thus, a collaborative explanatory 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 (Chi & VanLehn, 1991), and those who actively explore the system in order to solve particular problems. So, CXAI 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 its shortcomings? What are some suggestions and methods for working around the shortcomings? CXAI may help provide user-centric explanations that do not require algorithms, the creation of formal user models, or complex visualizations in order to provide important explanations. Furthermore, the explanations that are elicited may complement those produced by algorithmic approaches, providing a different level of information that may be made even more useful and actionable. This type of approach is well-established in social Q&A (SQA) communities but has yet to be applied and investigated in the XAI domain.

1.2 Overview of the Thesis

Next, in Chapter 2, I review previous literature mainly on collaborative work, and social Q&A. Although collaborative explanation systems have not been used in the XAI domain,

it has precedent in general collaborative systems referred to as social Q&A (Oh, 2018). Traditional SQA approaches include message boards, platforms such as Yahoo Answers, and programming help boards such as StackExchange. Although the CXAI shares properties with these, it is also intended to help users focus on explaining a particular AI system's behaviors. Previous research also tells us how to initiate problem-solving in a collaborative setting that includes web-based collaboration and motivate users to participate in the collaboration. The next chapter (Chapter 3) is dedicated to the human-centric development of a CXAI system and the introduction of a web-based system (AI Database Browser) that hosts results from an Image Classifier (see Mueller et al. (2020), who examined the performance of the system). This system will be used in the baseline condition in the experiments that will be conducted as a part of this thesis.

In Chapter 4, I report on a heuristic formative evaluation for the CXAI system to assess 'explanation goodness' with the help of the 'goodness criteria' (Hoffman, Mueller, et al., 2018). The goodness criteria are a set of principles that guide the development of explanations that can often be reasonably evaluated without relying on users, centering on concepts such as correctness, completeness, incrementalism, reversibility, and the like. Another way to assess explanation is through the assessment of qualitative measures like satisfaction, trust, etc. (Hoffman, Mueller, et al., 2018). Also, to assess a user's mental model and performance with the XAI system, it is necessary to do conduct tests of comprehension and performance. Chapters 5 and 6 discuss results from the experiments that tested users' comprehension, performance, and additional qualitative measures. The final chapter is the conclusion of the thesis through general discussion.

2 Literature Review

The purpose of this literature review is to summarize some of the research that forms the precursor and precedent for understanding collaborative explanations in learning, problem-solving, and AI. In this chapter, I will discuss what type of knowledge is generated during collaborative learning and how collaborative learning can help in problem-solving. Also, since it is possible to keep track of collaboration in a web-based setting, factors that affect users' attitudes and cognitive load in collaborative web learning are discussed in this chapter. Collaborative filtering (CF) (X. Su & Khoshgoftaar, 2009) provides precedent of using user-generated data to solve a slightly different problem in AI. Collaborative sensemaking in the collaborative effort helps in understanding multiple opinions. These topics are discussed in this chapter. I will also discuss the benefit of using a Social Q&A system as a backbone for a system that would help users to generate collaborative explanations and collaborative tutors in tutoring in this chapter.

2.1 Collaborative Learning

One important area of literature in collaborative explanations is research on collaborative learning. Collaborative learning has a broad meaning, ranging from learning done in a small group to learning supported by complex internet-based systems. It can be conducted as a pair or in a group, face-to-face or computer-mediated, synchronous, or asynchronous (Dillenbourg, 1999). Collaborative learning can be generally described as a situation in which particular forms of interaction among people are expected to occur, which would trigger learning. Working together while accomplishing a task is seen as a characteristic of

a powerful learning environment because it fosters the active construction of knowledge (Van Merriënboer & Paas, 2003). As collaboration is an effective learning approach, it should help users of an AI learn from each other and form a better understanding of how the AI works.

2.1.1 Shared Knowledge Base & Critical Thinking

An important subdomain of collaborative learning is research on how people share knowledge that can support critical thinking. Thalemann & Strube (2004) found that knowledge is about the initial situation and goals are well-defined for a problem, users perform better in problem-solving. Partners can integrate their knowledge. Wertsch (1986) contended that cooperative learning improves problem-solving strategies because the students are confronted with different interpretations of the given situation.

Jeong & Chi (2000) found that during the development of a shared knowledge base, the collaborators add more knowledge in the knowledge base than non-collaborating partners due to collaboration. This knowledge may act as a catalyst in problem-solving. Students also develop problem-solving skills by formulating their ideas, discussing them, receiving immediate feedback, and responding to questions and comments (Johnson, 1971; Peterson & Swing, 1985).

Collaborative learning also helps in developing higher-level thinking skills (Webb, 1982). Learners 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).

2.1.2 Web-based Learning

Another subdomain in collaborative learning is the use of web-based learning tools that help in group knowledge sharing (Koschmann, 1996). Different learning technologies such as the Knowledge Community and Inquiry Model (Slotta & Najafi, 2013) use Web 2.0 technologies where students explore a conceptual domain, express their ideas, and create a collective knowledge base that future users can use. 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). Liaw et al. (2008) also found five factors that positively influence users' attitudes towards collaborative web learning. The factors are system functions, system satisfaction, collaborative activities, learners' characteristics, and system acceptance. So, a web-based system that generates collaborative explanations needs to consider all these factors to be acceptable to users. In this section, I will review the literature on communication in web-based learning that will highlight its impications for a functioning collaborative environment for collaborative explantions.

2.1.2.1 Communication in Web-based Learning

Effective communication is needed for efficient collaboration. It is required to understand what type of communication will be effective in a web-based system. Message exchange among users in a web-based platform can offer certain advantages. In asynchronous communication, it is possible that messages can be read and answered without any hindrance and with ample time. Topics can be discussed temporally in parallel and

separated into topics making structured discussions for larger groups possible (Hron & Friedrich, 2003). Schwan et al. (2002) found mutual message exchange can take place anytime during asynchronous communication removing any need for turn-taking. To mitigate any unnecessary cognitive load that may occur from the exchanges supporting measures need to be introduced to avoid negative consequences for learning due to unnecessary cognitive load (Kirschner, 2002). This can be done by introducing supporting materials like sharing references related to a problem.

It is critical to be aware that there is no guarantee that the desired activities in web-based learning groups will occur. Some form of structuring may be helpful to guide users to the required tasks. Research has shown that implicit and explicit dialogue structuring showed greater orientation on the subject matter (Hron et al., 2000). Weinberger et al. (2002) also reported that scripted cooperation was beneficial for individual transfer, knowledge convergence, and participation in small groups.

Another way in which web-based collaborative learning happens is through inquiry learning. Inquiry learning has four basic features; generating hypotheses, collecting data, interpreting evidence, and drawing conclusions (Looi, 1998; Suthers, 1996; White & Frederiksen, 1998). A web-based collaborative explanatory system might allow users to implement all these processes, putting forward a hypothesis about an AI system that will lead to a collective collection of data, interpreting evidence, and generating a consensus.

Users may lack the motivation to make deep inquiries if the users cannot connect with the topic. So, the presence of sufficient motivation can initiate desired activities in web-based learning (see next section to learn what motivates users in a social Q&A platform). A web-

based system also allows users to revise their concepts and change their original thoughts that are possible in web-based learning (Chang et al., 2003).

Web-based Open Annotation Collaboration (Haslhofer et al., 2011) or Social Annotation (SA) (Kalir, 2020; Novak et al., 2012) is another approach to collaborative learning. Social Annotation enables collaboration when learners perceive and engage with texts as dialogical contexts. It encourages learners to share their subjective interpretations. The exposition of ideas happens during the annotation conversations. Contributors to a group inquiry develop associative connections, helping them to recognize multiple perspectives (Kalir & Garcia, 2019). Openly networked SA also motivates knowledge construction (Chen, 2019). Users in a collaborative environment become satisfied with their experience when they can create individual annotations and share their own annotations in a collaborative learning context. The influence of annotation on learning achievements becomes stronger with the use of the sharing mechanism (A. Y. Su et al., 2010).

One of the shortcomings of web-based annotation is that users need to navigate many text-based annotations despite users showing engagement in a variety of behaviors, including self-reflection, elaboration, internalization, and showing support while using social annotation tools (Gao, 2013).

2.2 Social Q&A (SQA)

SQA occurs in a social context in which people ask, answer, and rate content while interacting around it (Oh, 2018). SQA platforms need to be public, community-based, and reliant on natural language (Shah et al., 2009). SQA systems serve as public or community-

based resources and rely on natural language communication (Shah et al., 2009) rather than extensive algorithmic data or video. So, collaborative explanations about AIs could be generated in a specialized SQA platform that supports user groups for AIs. In order to succeed, however, users of an SQA platform need to be sufficiently motivated to interact with the collaborative system. A small community or team may be motivated to communicate intrinsically, but other SQA systems have incorporated specific features that encourage contributions. Responders' authority, shorter response time, and greater answer length are some critical features that positively associate with the peer-judged answer quality in an SQA site (Li et al., 2015). This section will discuss how to motivate people to make quality posts on a social Q&A platform.

2.2.1 Ensuring the Quality of Posts

With the growth in popularity of social networking sites, evaluating the quality of the information they contain has become increasingly important. The requirement for quality of answers can change for different types of questions. Older scholars in academic SQA sites tended to view verifiability as more important to the quality of answers to information-seeking questions than to discussion-seeking questions (Li et al., 2020). In an informal social context, users do engage in more informative conversations (does not worry about accuracy) but withhold information aiming not to hurt accuracy in a formal social context (Martín-Luengo et al., 2018). Mentored questions based on feedbacks also improve the quality of questions that can lead to better answers. This method is also satisfactory to novice users (Ford et al., 2018). Certain interaction acts such as upvote/downvote can also

be rewarding generally in terms of attaining higher perceived post quality in an SQA platform (Sin et al., 2018).

2.2.2 Motivation & Reputation

To ensure desired learning activities occur in a collaborative web-based environment, users need to be motivated which was mentioned previously in the Collaborative Learning section. So, an important subdomain for Social Q&A is motivation and reputation, this subdomain discusses how users can be motivated to use an SQA platform. SQA sites vet existing contributions and motivate future contributions by awarding points to users (Oh, 2018). SQA sites 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 also contributes his/her knowledge because of factors including the user's self-presentation, peer recognition, incentive, etc. (Jin et al., 2015; Khansa et al., 2015). The best answers in an SQA platform are correlated with the consistent participation of users. A structure based on points can further motivate participation (Nam et al., 2009). One of the ways to motivate users to answer questions and increase the reputation of the answerer is to introduce a 'bounty system' such as that in Stack Overflow (Zhou et al., 2020). In an organizational environment, points through a bounty system can drive users to share explanations about an AI system. Number of up/down votes on a statement indicates quality of answer (Jeon & Rieh, 2013) that can also increase or decrease reputation of the answerer.

2.3 Collaborative Filtering (CF)

Though goals for Collaborative Filtering or CF are different from the goals of a system that generates collaborative explanation, the CF approach has some important features that may help in building an effective collaborative explanatory system. Both CF and the collaborative explanation in AI have had a similar journey till now. Typical recommender systems and current explanatory systems depend on algorithms to give recommendations or explanations. Both CF and collaborative explanations depend on humans to make recommendations or generate explanations.

CF uses a database of preferences for items by users to predict additional topics or products a user might like (X. Su & Khoshgoftaar, 2009). Database preferences for a user are built based on the likes/dislikes of other like-minded users. This provides a collective idea of user preferences in a system. This approach is used for building recommendation systems that rely heavily on correlations among user preferences instead of complex taxonomies or AI analysis of the products, movies, music, or applications being recommended.

One of the CF techniques is Memory-based CF which is mainly dependent on users' ratings. This method is easy to implement and shows good performance for dense datasets (X. Su & Khoshgoftaar, 2009).

Tapestry (Goldberg et al., 1992) is one of the earliest electronic document filtering systems that use CF. Users were encouraged to annotate documents somewhat like tagging, and these annotations were then used for filtering. The researchers identified two types of users, eager and casual users. Eager users annotated the documents, and casual users waited for

eager users' annotation to do filtering. Another approach of CF is the Search-Based Method (Linden et al., 2003) where it shows items that have related keywords or subjects to an item.

I expect users to help each other expedite the search process for explanations regarding an AI's trait on any collaborative explanatory system. This can be done through tagging, voting, or categorizing items under some specific topics. So, adding features like keywords or tags in the collaborative explanatory system may help users to expedite the search process.

2.4 Collaborative Sensemaking

Sensemaking is the process of understanding complex situations or phenomena. While generating collaborative explanations, people will be learning and making sense of an AI system. In collective/team sensemaking, groups of individuals collaborate to develop understanding at both the individual and collective levels (G. Klein et al., 2010). Collective Sensemaking can be used to synthesize vast amounts of information and opinions. Mamykina et al. (2015) studied users of an online diabetes community, TuDiabetes. The researchers found that the users often construct shared meaning through deep discussions, back and forth negotiation of perspectives, and resolution of conflicts in opinions. The users also expressed a multiplicity of opinions rather than confirming a consensus. Zagalsky et al. (2016) also found that users co-create knowledge in collaborative sensemaking in their study on the R community.

2.5 Collaborative Problem Solving

While creating collaborative explanations, users will not learn from each other about an AI system through user-crafted explanations about the AI system only, they first need to figure out unknown traits of the AI system together. Problem-solving depends not only on making sense of the machine's result but also on the division of labor in the group. This section will discuss the effective way to initiate problem-solving in a collaborative setting that will guide for creating an environment for understanding the AI system. In a web search task, for initial, and synchronous search, a chat-centric view was preferred by 67% of participants in the CoSense tool because they can go back and look for solutions to a problem (Paul & Morris, 2009). Though chat-centric communication is not asynchronous communication and does not offer the benefits of asynchronous communication discussed in the 'Communication in Web-based Learning' section of this thesis, a chat-centric option will still be useful for forming explanations about an aspect of an AI system that is not understood. Chat-centric work helps in keeping track of what decisions are made in the group and how each member is performing in the task of problem-solving. In the initial stage of collaboration, where problem-solving has not started yet, specific questions can be useful to initiate problem-solving. Questions can be divided into several levels to tap different levels of knowledge. Four levels can be 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). Similar trigger questions have been examined in the scope of XAI (Mueller et al., 2019).

Collaborative problem-solving tasks include content-free and content-dependent types (Care et al., 2015). Content-free tasks depend on inductive and deductive thinking skills. Content-dependent tasks allow users to draw on knowledge gained through traditional learning areas or subjects. To ensure content-dependent tasks occur in a collaborative platform, additional references can be added to the problems so the users can draw the knowledge from these references. This thesis will use specific types of questions that will drive users to share explanations about an AI system in a collaborative setting. Also, it will use a reference system to familiarize users with an aspect of an AI besides an explanation that will draw opinions from other users.

2.5.1 Questions to Generate Explanations

As we have seen earlier, for initiating problem-solving, questions can be an effective tool. With questions, we can generate explanations also. According to Gruber (1991), knowledge acquisition systems can be designed to ask why-questions in the form of justifications besides asking 'what' – style questions. This way, it will be easy to explain a particular action, such as a decision to take some action or choose an appropriate alternative in a given situation. Questions have been used in the explanatory process of the AI system in the past. AQUA system (Ram, 1993) used questions to generate explanations and fill up knowledge gaps. Curiosity can motivate users to ask for explanations about an AI system through questioning (Hoffman, Mueller, et al., 2018). Liao et al. (2020) created an algorithm-informed XAI question bank with prototypical questions that users may ask to

understand AI systems. This approach of using questions for problem-solving may also be useful for understanding the "black-box" nature of an AI system.

2.6 Collaborative Tutoring

Intelligent Tutoring Systems have been used effectively for scientific problem solving (Friedland et al., 2004), electronic circuit design (Brown & Burton, 1978), propulsion engineering (Stevens & Roberts, 1983), etc. Intelligent Tutoring Systems aims to promote adaptive interaction between the learner and the content (George et al., 2016) through Socratic dialog. The main challenge has been to create intelligent tutoring systems that are user-adaptive.

Critics may argue that explaining an AI system cannot be done collaboratively by humans; even explanations given by humans will not be helpful if they are wrong. But research has shown that people can also learn from errors if they recognize them (Chi et al., 2001; VanLehn et al., 2003).

Human collaboration is also highly effective in promoting learning. In one study, a pair of learners tutored one another. A third person observed collaborative tutoring. The results showed that observing collaborative tutoring can help a learner as much as being directly tutored in a tutoring dialog (Chi et al., 2008). It was the collaboration that was helpful. Tutoring effectiveness does not depend only on tutors' pedagogical skills but also on substantive construction resulting from interactions between the tutors and the students. A joint effort between tutors and students can improve learning (Chi et al., 2001). The ICAP (Interactive>Constructive>Active>Passive) Framework tells us that learning is better when

human-human interaction is present through dialoguing (Chi & Wylie, 2014). In a collaborative matter, it may be possible for humans to learn about an AI system better from other humans than learning from an intelligent system.

The chapter mainly put forward the elements that might help build a non-algorithmic collaborative explanatory system. Questions that can help to initiate problem-solving, keywords for collaborative filtering, and motivation to answer questions are some of the findings that can be incorporated into the system. This section also discussed if it is possible to explain AIs through collaborative tutoring.

Having disclosed these design requirements, I next proceeded to design a collaborative explanatory system and then "populate" it with data so that it might be evaluated.

3 Basic Design of the Collaborative System (CXAI System)

Based on the literature review, I developed a web-based novel explanatory system (CXAI Tool/CXAI System) similar to a social QA platform but modified for explaining AI systems. The system has standard features of a general social QA platform (like StackOverflow) where users can associate keyword(s) to their posts, a system of bounty to engage users in the platform; 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 1 for topics) for explanations that have been revealed in the research on the importance of users' goals and needs regarding explanations (Mueller et al., 2019). Once users select one or more topics to frame their answers, it would serve as metadata to contextualize the user's notes and the responses from other users. This might support other users' search through the collaborative system.

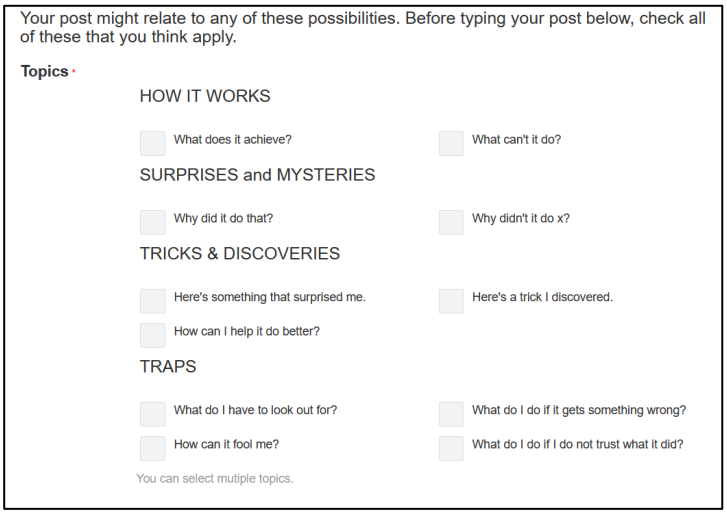


Figure 1. Topics as ‘triggers’

This system's users can also add reference(s) to their posts about the AI system that they are using so that other users can understand the posts with the help of the reference(s).

3.1 Evaluation of the CXAI System

The system was evaluated for usability issues using Nielsen's 10 Usability Heuristics for User Interface Design (Nielsen, 1994). This gave quick feedback on the design. After solving the usability issues that were found from heuristic evaluation, the system was loaded with dummy data for evaluation by two test users. This part of the evaluation was focused on searching entries by users. The test users were tasked to search a few dummy cases that were present in the collaborative system. When they did the searches, they were also asked to use the think-aloud protocol (Jääskeläinen, 2010). From this part of the evaluation, it was apparent to the test facilitator; the test users were ignoring the comments of the posts, only looking for answers for the cases in the posts. So, one of the major modifications from this evaluation process is to place the relevant comments with the relevant posts after a search (see Figure 2).

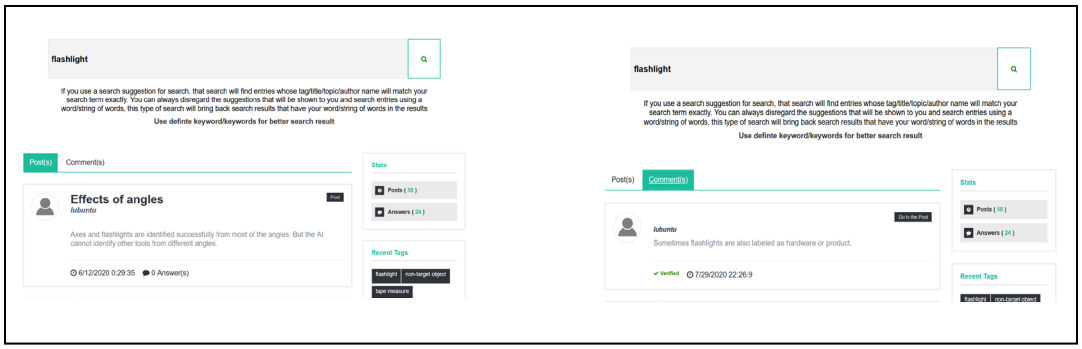


Figure 2. A modification of the system after the think-aloud protocol

3.2 Populating the CXAI System

After the system upgrade, dummy data from the system was removed. An image classifier was selected for evaluating the collaborative system. To make the image classifier easy to use during data entry in the collaborative system, a new system was used (Mueller et al., 2020), I call it the 'AI Database Browser'.

3.2.1 AI Database Browser

Images of different types of tools were selected for image classification. Each tool was photographed in ten different conditions or transformations, for example, a tool was photographed inside a leafy frame, the same tool was transformed into a sketch, etc. These images of the tools were classified using the image classifier. For each image, a table was created with the results for that image. The table shows the labels the AI provides, along with its confidence score in that label and the correctness of the classification.

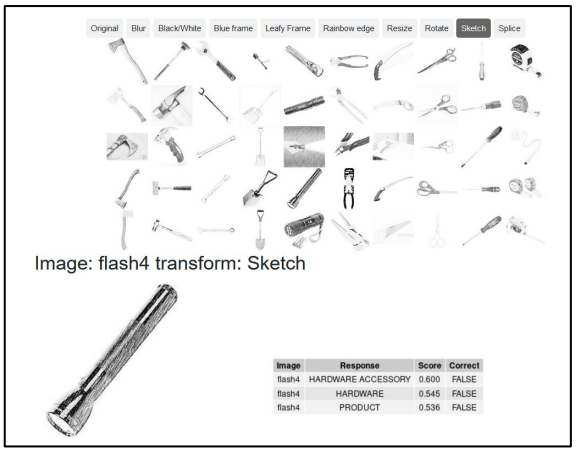


Figure 3. AI Database Browser showing a sketched flashlight with the classification results

Each image with its respective classification results was hosted on a website (see Figure 3 to see the interface of the AI Database Browser). Using this system, a user can browse the

images, see each image's results, and determine how the image classifier performs for each condition/transformation.

3.2.2 Data Entry & Verification

The AI Database Browser was handed to a group of participants with the CXAI system (no dummy data was present in the CXAI system during this process). The group of participants browsed the AI Database Browser and posted their observations of the AI system in the CXAI system. Their observations mainly include the performance of the AI system in different transformations, tricks to get correct image identification from the AI system, shortcomings of the AI system. They also collaboratively discuss the posted observations via comments. This way, explanations may have been refined by adding knowledge to main observations. Another group of participants were given the same AI Database Browser and asked to do the same procedure. But this time, the latter group did not make any entry to the CXAI system, rather listed their observations in an excel sheet. The purpose of this procedure is to see if both the groups independently report the same findings on the AI system, and after comparing the two sets of records, it was found that except for one or two occasions, the two sets of record match. Also, the two groups reported the same type of shortcomings about the AI system. This validates the claim that the CXAI system has the necessary entries about the AI system. Figure 4 briefly presents the development of the CXAI system through visual representation.

Figure 4. Development of the Collaborative System (CXAI Tool/System)

4 The Goodness Criteria

Hoffman, Klein, et al. (2018) and Hoffman, Mueller, et al. (2018) (see also Mueller et al., 2021) described a comprehensive measurement approach for assessing explanations in the context of AI systems. This included (1) judgements explanation 'goodness'; (2) assessment of user's mental models; (3) judgements of qualitative measures of trust, satisfaction, and reliance; and (4) evaluation of human-AI task performance. Although many of these measures have been widely employed in the XAI domain, the first category, "explanation goodness" has received less adoption and investigation. The goodness criteria are a set of principles that guide the development of explanations, that can often be reasonably evaluated without relying on users, centering on concepts such as correctness, completeness, incrementalism, reversibility, and the like.

Principles of 'goodness criteria' can affect an XAI system in many ways. 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 could impact different aspects of satisfaction, and so it is important for heuristic evaluation. Related to this is the **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 the measures of readability can provide insight into this criterion. Lastly, an XAI system should be a **knowledge** base to future users.

Although 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 remains a gold standard evaluation, 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 thesis will evaluate the explanations generated in the CXAI system with this set of new criteria to determine the system's strengths as an XAI system and provide an example evaluation approach for examining future XAI systems.

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 correctness (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 Correctness Criterion

One might expect that novice users will provide explanatory statements that are often incorrect. Consequently, we coded the correctness 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 correctness, 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 correctness (e.g., removing

opinion). Chunks are explanations from the CXAI system that were broken down based on the agreement between two raters if they thought a single explanation is talking about more than one aspect of the AI. Foils were statements that are true for another Image Classifier that is not true for the Image Classifier on which the explanations of the CXAI system/tool are based. The purpose of the foils was to ensure that the coders could accurately distinguish between correct and incorrect statements about the target classifier.

The raters achieved a moderate level of agreement on the cases (weighted κ=0.76). Of the 79 target statements (see Table 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 correctness, and thus that the target explanations achieved a high level of correctness.

Table 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	**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. Here scope is defined 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, 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 2.

Table 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 a strong agreement between them according to the κ value, 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 & Dey (2009), which identifies five basic questions explanations answer: What, Why, Why Not, What If, and How To. Two independent coders coded 95 original chunks to evaluate explanation type 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 $\kappa=0.76$. Their coding results can be summarized in Table 3, which demonstrates that the CXAI explanations mostly answered 'what' questions.

Table 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 Table 4). 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 (five basic questions) to the user-specified trigger codes (see Table 5). 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 & Dey (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/required in current XAI systems (Tosun et al., 2020; Wick & Thompson, 1992). However, the raters identified substantial numbers of explanations as answering 'why questions' that were coded as 'what' explanations (see Table 6 for examples). This may be because the explanations were cued by asking the 'why' question but did not provide a 'why' answer.

Table 4. Triggers in the XAI system (CXAI)

Type	**Triggers**	
How it works?	What does it achieve?	What can't it do?

Surprises and Mysteries	Why did it do that?	Why didn't it do x?
Tricks & Discoveries	Here's something that surprised me.	Here's a trick I discovered.
	How can I help it do better?	
Traps	What do I have to look out for?	What do I do if it gets something wrong?
	How can it fool me?	What do I do if I do not trust what it did?

Table 5. 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

Table 6. Examples for explanation type

Explanations	**Lim & Dey (2009)**	**Trigger**
the black and white and sketch versions are similar. They generally provide different sets of responses. They are similar, in that they tend to focus on broader categories, or some things like 'body jewelry', but the same image gives different outputs. This is somewhat surprising.	What?	Why didn't it do x?
I see a tendency of the system to incorrectly classify the tools as plants or objects relating to plants when the frame intrudes on the integrity of the image or is very close to it. I'm not sure if the AI is able to split up the image (e.g. non intrusive frame separate from the actual object) or not.	What?	Why didn't it do x?
Regardless of whether the A.I. successfully identifies target object with rainbow edge transformation, in many cases, the A.I. will also recognize the images as computerized image (computer wallpaper) with an x-ray look. Not sure if this feature will be useful for anything.	What?	Why did it do that?
When images are framed with a blue frame, the AI does not always get it right. But it usually gets it wrong only when the original	What?	Why didn't it do x?

image was wrong. In some cases, the same image was labelled correctly under the blue frame even when it was in error in the original. It seems like the blue frame does not impair the AI consistently		

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).

4.5.1 Compared with other explanatory text

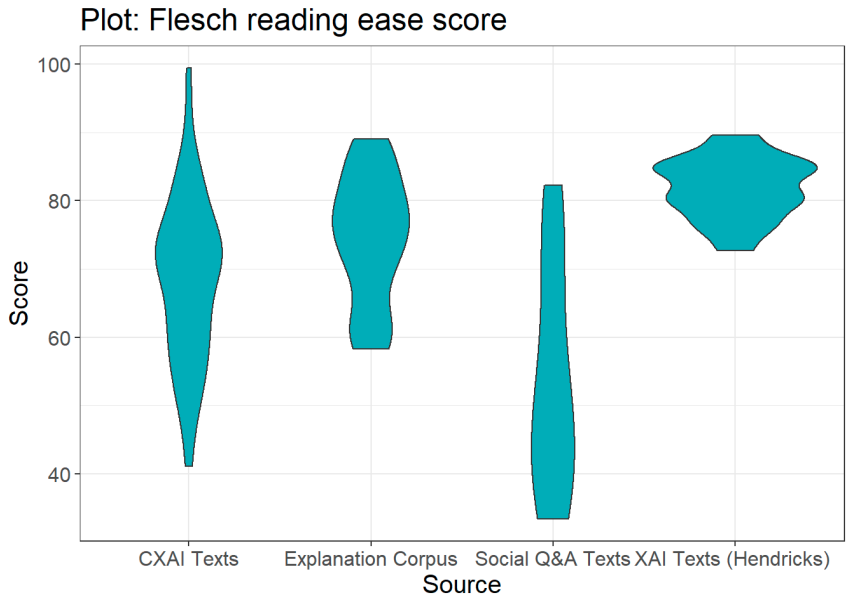
To compare the simplicity of the CXAI explanations with other explanations, we examined a corpus of explanations collected from the internet, popular press, and other sources (G. Klein et al., 2019) about general topics. These explanations covered many kinds of complex systems outside of the AI domain. These statements produced a mean Flesch–Kincaid grade level score of 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 social Q&A texts 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) as texts were generated through algorithm(s) describing different pictures. 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 5). 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 on-line message boards, slightly more complex than AI-generated text explanations.



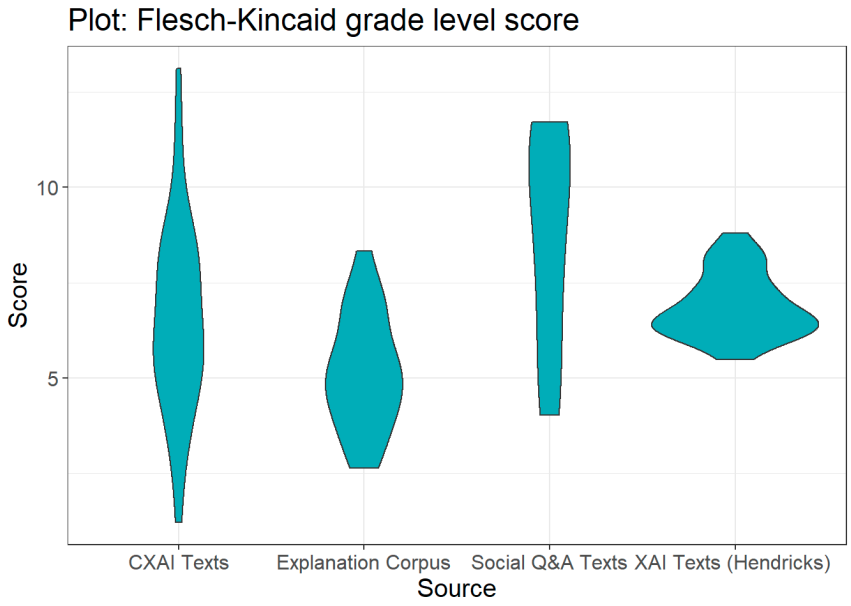


Figure 5. Distributions of Flesch Reading ease scores (top panel) and Flesch-Kincaid Grade Level Scores (bottom panel) for CXAI explanations in comparison to three other explanation corpora.

4.6 Discussion

This analysis demonstrates how a heuristic evaluation of the contents 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.

5 Test of Comprehension and Performance – User Study 1/Experiment 1

According to Hoffman, Mueller, et al. (2018), an XAI system should enable users of an AI system to show better performance with the AI system. So, a test of performance is required to assess a novel XAI system. A novel XAI system is also assessed by the test of comprehension that will show if the users understand the AI system through the XAI system. Experiment 1 is designed to address both the issues regarding performance and comprehension for an XAI system. This 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. This chapter is part of Experiment 1.

5.1 Method

5.1.1 Participants

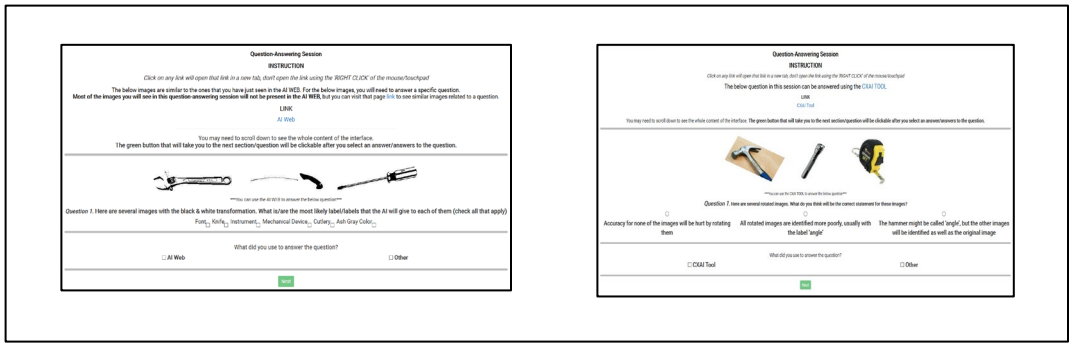
69 undergraduate students from Michigan Technological University (MTU) participated the Experiment 1 in a credit-based compensation structure. Through a video tutorial, these participants were trained to use at least the AI Database Browser or the CXAI system.

5.1.2 Procedure

For Experiment 1, a set of questions (10) about the AI system was created. The questions represent all the transformations of the AI Database Browser (see Figure 6 for examples of

questions). The questions asked the participants how the AI will 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. Except for two questions, new pictures of the tools were used related to the tools present in the AI Database Browser. These 10 questions will be used to test the performance and comprehension of the novel XAI system. I am testing this novel system in an ideal condition where all the answers to the questions can be found in the AI Database Browser and the CXAI system. Users of the CXAI system can answer each question with the help of explanations generated through collaboration when the CXAI was populated (see section 3.2 of Chapter 3 for more on populating the CXAI system). In the between-subject design, participants used AI Database Browser and the CXAI tool/system.

In both conditions, a participant can self-report if they used a particular system or used other means (for example, guessing) to answer a question. After agreeing to the consent form, and answering a few demographic questions, a participant was trained on a particular system (AI Database Browser/CXAI system) based on the system the participant was assigned to, with the help of a video on the system. After that, the participants answered the questions without time constraints.



(a) (b)

Figure 6. (a) A question in the control condition, (b) A question in the experimental condition

At the end of the Question-Answering session, the participant answered two open-ended questions that asked them what made the session easy and difficult for them.

5.2 Results

5.2.1 Overall Accuracy & Time

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; $t(66.67) = -2.21$, $p = 0.03$; $d = 0.56$.; see Figure 7 to see the graph on accuracy).

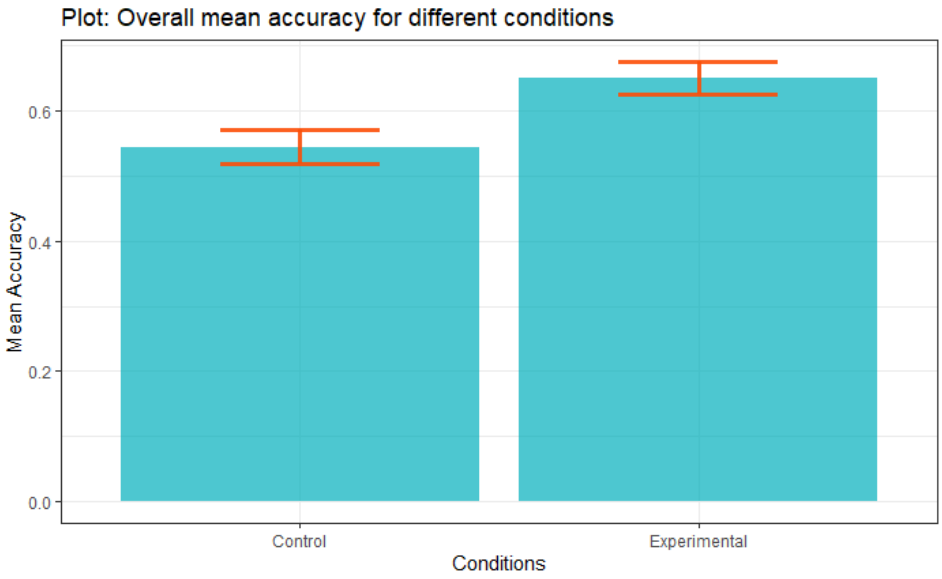


Figure 7. Overall mean accuracy for the conditions

It is also useful to examine the time needed to answer the questions. Figure 8 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: ($D = 0.13$, $p = 0.86$). Though these results do not support our hypotheses completely because CXAI system users did not accurately answer the questions faster than the other system. But the users of the CXAI system took a similar amount of time (they were not slower) to the users of the AI Database Browser to achieve higher accuracy.

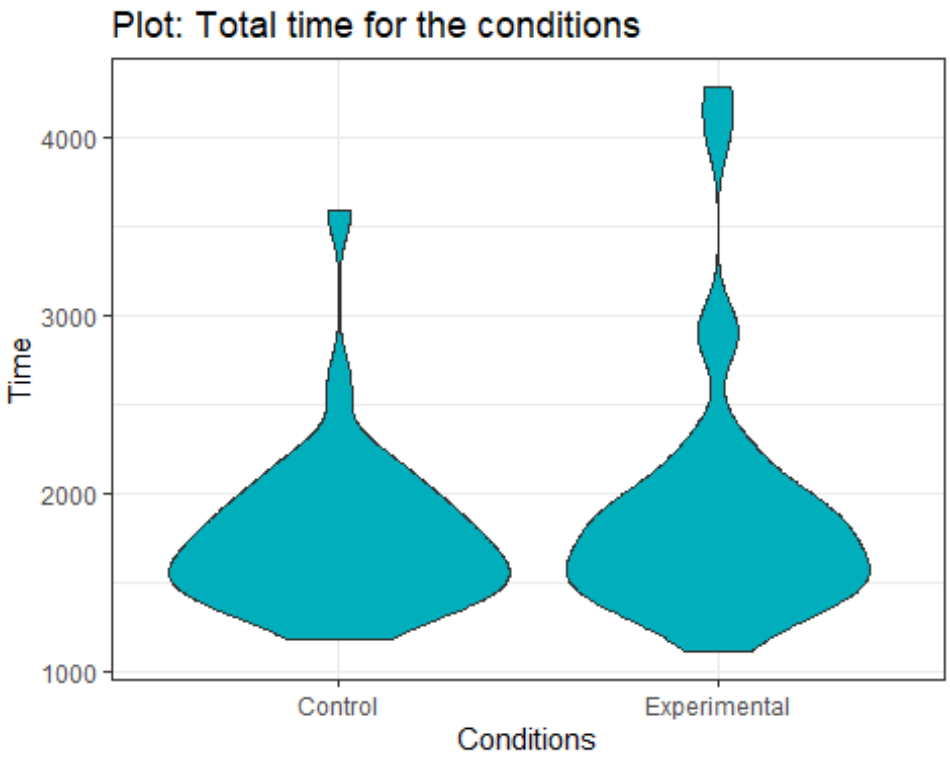


Figure 8. Total times for the conditions (distribution similar)

5.2.1.1 Accuracy during System Use

Now, the questions that were answered using only one of the systems (AI Database Browser or CXAI System) were taken into consideration. There are roughly 30 records per question in each condition. So, we got an even distribution of records for the questions in the conditions. 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 7). 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 ($X2(2) = 641.74$, $p < 0.001$.) This shows that users in the experimental condition tended to trade 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 a system, 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 7. Mean Accuracy for the system use/unuse

System	System Used	Mean Accuracy
AI Database Browser	Yes (n = 324)	0.55
AI Database Browser	No (n = 16)	0.25
CXAI System	Yes (n = 301)	0.73
CXAI System	No (n = 49)	0.26

5.3 Discussion

The user study results reported here show that collaborative explanations can be helpful, insofar as they help produce accurate answers to questions about the AI system while not taking substantially longer to answer. 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 Experiment 1, participants self-reported that they guessed about 3 times more often when using the CXAI system than when browsing the AI Database Browser directly. This may stem from the ease with which some questions could be investigated using the AI Database Browser, or the challenge of finding relevant CXAI entries related to particular questions. The result also indicates that people can correctly

answer questions about an AI system using the explanations generated by relatively novice users of the system.

6 Assessment of Qualitative Measures – User Study 2/Experiment 2

User behavior often changes depending on the feeling of satisfaction while using an IS - information system (Gatian, 1994). Many researchers agree that emphasis in IS research has been shifted from efficiency measures toward effectiveness measures such as user satisfaction (Sink et al., 1984). It has been denoted as an important surrogate measure of information systems success (Bailey & Pearson, 1983; Baroudi et al., 1986; Benson, 1983; Ives et al., 1983). A reason for this shift is because of the psychological expectancy theory that says attitudes (i.e., satisfaction) are linked to behavior (i.e., productivity) (Fishbein, 1967; Fishbein & Ajzen, 1977). Efficiency and decision-making performance are both correlated to user satisfaction for the users who directly use a system (Gatian, 1994). One of the measures to assess explanations from an XAI system is measuring the satisfaction level of the explanations (Hoffman, Mueller, et al., 2018). Presumably, users might not notice improvements in accuracy, and so subjective measures might be important for predicting adoption of the tool. Furthermore, Experiment 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, I assessed explanations from the collaborative platform using different qualitative measures. This chapter will test the satisfaction level for the explanations from the CXAI System using some key attributes (satisfaction, sufficiency, completeness, trust) from the 'Explanation Satisfaction Scale' (Hoffman, Mueller, et al., 2018). This chapter is part of Experiment 2.

6.1 Method

6.1.1 Participants

43 undergraduate students from Michigan Technological University (MTU) participated in Experiment 2 a credit-based compensation structure. These participants were briefed on the AI Database Browser or the CXAI system.

6.1.2 Procedure

The participants were given a made-up scenario where a participant was attached to a Hardware Store where two explanatory systems are used (AI Database Browser and CXAI system) to explain Hardware Store AI's decision to customers. Unlike Experiment 1, the experimental design was within-participant, so that each participant used both the CXAI and AI Database Browser. The participants were given 8 questions regarding different instances, transformations, or tools (see Mueller et al. (2020)). There were two counterbalancing conditions (Condition 1 and Condition 2 – see Figure 11). In Condition 1, a participant answered odd number questions using AI Database Browser, and even number questions were answered using the CXAI system and this was vice-versa for a participant in Condition 2. A sample of explanations regarding the instance, tool, or transformation was attached from the CXAI System or AI Database Browser for each question. The three best examples determined by a group of researchers from Mueller/Veinott Lab at MTU 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 System regarding the instance, tool, or transformation were given for the CXAI System 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 Hoffman, Mueller, et al. (2018), where a 7 denotes 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.

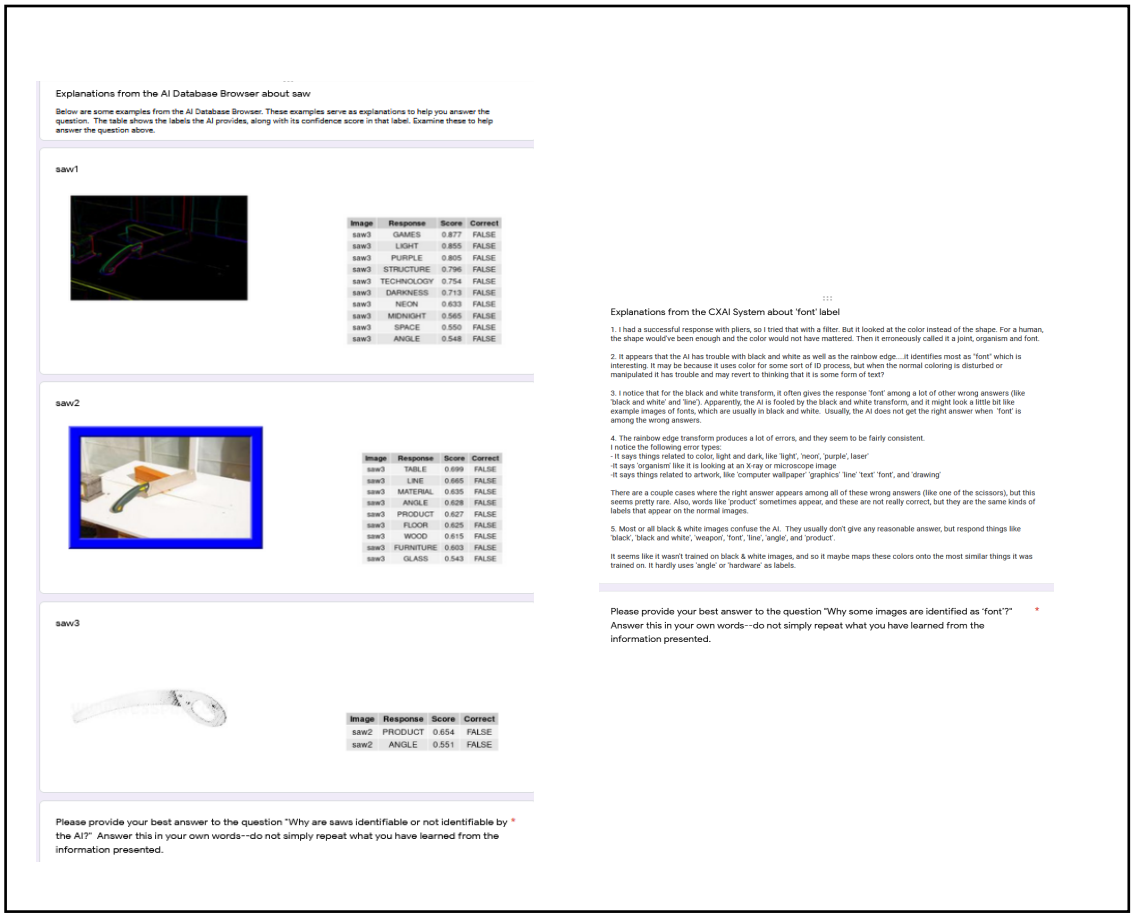


Figure 9. Sample of explanations from the systems; Left panal shows explanations from the AI Database Browser and Right panal shows explanations from the CXAI System

6.2 Results

For all the attributes (satisfaction, sufficiency, completeness, trust), CXAI system generated higher ratings than AI Database Browser. 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: $t(86) = -4.17$, $p < 0.001$; $d = 0.32$.

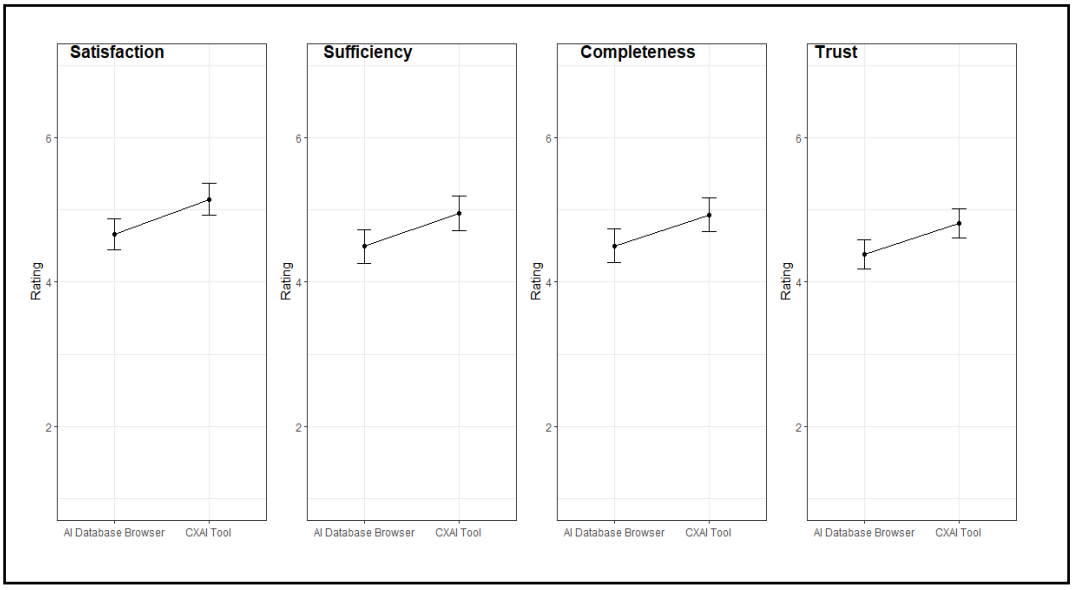


Figure 10. Comparison of the two systems on the attributes (satisfaction, sufficiency, completeness, trust)

6.3 Discussion

Though there may have been a trading-off accuracy for effort (Liesefeld & Janczyk, 2019) in User Study 1, participants rated explanations from the CXAI system as more satisfying, sufficient, complete, and trustworthy in comparison to example-based explanations obtained by browsing the database itself.

7 General Discussion

In this thesis, I have described the motivations and iterative design processes for developing the CXAI system. The thesis evaluated both the system and the content of the system through heuristics evaluation and user studies. In this thesis, for the first time, heuristic evaluation was done on the system itself and the content (through 'goodness criteria') for an XAI system. In user study 1, the result showed the users learned from the system that was reflected in their performance in the first user study. The second user study confirmed that they were satisfied with the explanations that they received from the CXAI system. CXAI system removes any requirement of building user models by letting users be the "intelligent tutors" for other users. This way, dependency on algorithmic explanation-based systems can be reduced that is dependent on AI's architecture for any change (Das & Rad, 2020). Overall, this novel non-algorithmic approach satisfied all the XAI measures (Hoffman, Mueller, et al., 2018) for evaluating a new XAI system.

7.1 Limitations

This thesis did the assessment of comprehension & performance and qualitative measures separately. This precluded me to assess users' reactions if an answer is absent for a knowledge question in the CXAI system. This arises another limitation, the user studies were conducted in an ideal scenario where the answer to the knowledge questions can be found in both the systems (AI Database Browser and CXAI System). It is also uncertain if the explanation forms will be different for a different group of users. As it was the first stage for experimenting with explanations generated through this type of non-algorithmic process, testing it in an ideal condition is justified.

7.2 Future Directions

The next stage for the CXAI system is to implement it in a different user group like a group of radiologists in a less ideal condition. Also, required modifications to the system are needed to make it standalone. This way, it will be easy to measure how the system performs in an organizational setting. This will help me to understand if the CXAI system generates different types of explanations for different user groups. This will also help me to understand if many of the features of the social Q&A will help in enriching collaboration.

8 References

Alam, L. (2020). Investigating the Impact of Explanation on Repairing Trust in Ai Diagnostic Systems for Re-Diagnosis.

Bailey, J. E., & Pearson, S. W. (1983). Development of a tool for measuring and analyzing computer user satisfaction. Management Science, 29(5), 530–545.

Baroudi, J. J., Olson, M. H., & Ives, B. (1986). An empirical study of the impact of user involvement on system usage and information satisfaction. Communications of the ACM, 29(3), 232–238.

Benson, D. H. (1983). A field study of end user computing: Findings and issues. Mis Quarterly, 35–45.

Brown, J. S., & Burton, R. R. (1978). A paradigmatic example of an artificially intelligent instructional system. International Journal of Man-Machine Studies, 10(3), 323–339.

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.

Chang, K.-E., Sung, Y.-T., & Lee, C.-L. (2003). Web-based collaborative inquiry learning. Journal of Computer Assisted Learning, 19(1), 56–69.

Chen, B. (2019). Designing for networked collaborative discourse: An UnLMS approach. TechTrends, 63(2), 194–201.

Chi, M. T., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. Cognitive Science, 32(2), 301–341.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. Cognitive Science, 25(4), 471–533.

Chi, M. T., & VanLehn, K. A. (1991). The content of physics self-explanations. The Journal of the Learning Sciences, 1(1), 69–105.

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?

Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. ArXiv Preprint ArXiv:1702.08608. https://arxiv.org/abs/1702.08608

Fishbein, M. (1967). Attitude and the prediction of behavior. Readings in Attitude Theory and Measurement.

Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research.

Flesch, R. (1946). The art of plain talk.

Ford, D., Lustig, K., Banks, J., & Parnin, C. (2018). " We Don't Do That Here" How Collaborative Editing with Mentors Improves Engagement in Social Q&A Communities. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12.

Friedland, N. S., Allen, P. G., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., & Staab, S. (2004). Project halo: Towards a digital aristotle. AI Magazine, 25(4), 29–29.

Gao, F. (2013). A case study of using a social annotation tool to support collaboratively learning. The Internet and Higher Education, 17, 76–83.

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

George, S., Michel, C., & Ollagnier-Beldame, M. (2016). Favouring reflexivity in technology-enhanced learning systems: Towards smart uses of traces. Interactive Learning Environments, 24(7), 1389–1407.

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

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35(12), 61–70.

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

Gruber, T. (1991). Learning why by being told what: Interactive acquisition of justifications. IEEE Expert, 6(4), 65–75.

Haslhofer, B., Simon, R., Sanderson, R., & Van de Sompel, H. (2011). The open annotation collaboration (OAC) model. 2011 Workshop on Multimedia on the Web, 5–9.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2016). Generating visual explanations. European Conference on Computer Vision, 3–19. http://link.springer.com/chapter/10.1007/978-3-319-46493-0_1

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.

Hot Questions—Stack Exchange. (n.d.). Retrieved March 6, 2021, from https://stackexchange.com/

Hron, A., & Friedrich, H. F. (2003). A review of web-based collaborative learning: Factors beyond technology. Journal of Computer Assisted Learning, 19(1), 70–79.

Hron, A., Hesse, F. W., Cress, U., & Giovis, C. (2000). Implicit and explicit dialogue structuring in virtual learning groups. British Journal of Educational Psychology, 70(1), 53–64.

Huang, G., Fang, C. H., Agarwal, N., Bhagat, N., Eloy, J. A., & Langer, P. D. (2015). Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmology, 133(4), 449–454.

Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. Communications of the ACM, 26(10), 785–793.

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

Jeon, G. Y., & Rieh, S. Y. (2013). The value of social search: Seeking collective personal experience in social Q&A. Proceedings of the American Society for Information Science and Technology, 50(1), 1–10.

Jeong, H., & Chi, M. T. (2000). Does collaborative learning lead to the construction of common knowledge? Proceedings of the Annual Meeting of the Cognitive Science Society, 22(22).

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.

Kalir, J. H. (2020). Social annotation enabling collaboration for open learning. Distance Education, 1–16.

Kalir, J. H., & Garcia, A. (2019). Civic writing on digital walls. Journal of Literacy Research, 51(4), 420–443.

Khansa, L., Ma, X., Liginlal, D., & Kim, S. S. (2015). Understanding Members' Active Participation in Online Question-and-Answer Communities: A Theory and Empirical Analysis. Journal of Management Information Systems, 32(2), 162–203. https://doi.org/10.1080/07421222.2015.1063293

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch.

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. Elsevier.

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

Klein, G., Hoffman, R., & Mueller, S. (2019). Naturalistic Psychological Model of Explanatory Reasoning: How people explain things to others and to themselves. International Conference on Naturalistic Decision Making.

Klein, G., Wiggins, S., & Dominguez, C. O. (2010). Team sensemaking. Theoretical Issues in Ergonomics Science, 11(4), 304–320.

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

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

Li, L., He, D., Jeng, W., Goodwin, S., & Zhang, C. (2015). Answer quality characteristics and prediction on an academic Q&A Site: A case study on ResearchGate. Proceedings of the 24th International Conference on World Wide Web, 1453–1458.

Li, L., Zhang, C., & He, D. (2020). Factors influencing the importance of criteria for judging answer quality on academic social Q&A platforms. Aslib Journal of Information Management.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–15.

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.

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

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. Proceedings of the 11th International Conference on Ubiquitous Computing, 195–204.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76–80.

Looi, C. (1998). Interactive learning environments for promoting inquiry learning. Journal of Educational Technology Systems, 27(1), 3–22.

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.

Mamykina, L., Nakikj, D., & Elhadad, N. (2015). Collective sensemaking in online health forums. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 3217–3226.

Martín-Luengo, B., Shtyrov, Y., Luna, K., & Myachykov, A. (2018). Different answers to different audiences: Effects of social context on the accuracy-informativeness trade-off. Memory, 26(7), 993–1007.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282.

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.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 152–158.

Novak, E., Razzouk, R., & Johnson, T. E. (2012). The educational use of social annotation tools in higher education: A literature review. The Internet and Higher Education, 15(1), 39–49.

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

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. ArXiv Preprint ArXiv:1907.12652.

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.

Ram, A. (1993). Indexing, elaboration and refinement: Incremental learning of explanatory cases. In Case-Based Learning (pp. 7–54). Springer.

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

Schmidt, D. (2020). sylcount: Syllable Counting and Readability Measurements (0.2-2) [Computer software]. https://CRAN.R-project.org/package=sylcount

Schwan, S., Straub, D., & Hesse, F. W. (2002). Information management and learning in computer conferences: Coping with irrelevant and unconnected messages. Instructional Science, 30(4), 269–289.

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

Sin, S.-C. J., Lee, C. S., & Chen, X. (2018). Rewarding, But Not for Everyone: Interaction Acts and Perceived Post Quality on Social Q&A Sites. International Conference on Asian Digital Libraries, 136–141.

Sink, D. S., Tuttle, T. C., & DeVries, S. J. (1984). Productivity measurement and evaluation: What is available? National Productivity Review, 3(3), 265–287.

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.

Stevens, A., & Roberts, B. (1983). Quantitative and qualitative simulation in computer based training. Journal of Computer-Based Instruction, 10(1), 16–19.

Su, A. Y., Yang, S. J., Hwang, W.-Y., & Zhang, J. (2010). A Web 2.0-based collaborative annotation system for enhancing knowledge sharing in collaborative learning environments. Computers & Education, 55(2), 752–766.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009.

Suthers, D. (1996). Distributed tools for collaborative learning and coached apprenticeship approaches to critical inquiry. ITS'96 System Demonstrations.

Swartout, W. R., & Moore, J. D. (1993). Explanation in second generation expert systems. Second Generation Expert Systems, 543, 585.

Thalemann, S., & Strube, G. (2004). Shared knowledge in collaborative problem solving: Acquisition and effects. Proceedings of the Annual Meeting of the Cognitive Science Society, 26(26).

Tosun, A. B., Pullara, F., Becich, M. J., Taylor, D., Fine, J. L., & Chennubhotla, S. C. (2020). Explainable AI (xAI) for anatomic pathology. Advances in Anatomic Pathology, 27(4), 241–250.

Van Merriënboer, J. J., & Paas, F. (2003). Powerful learning and the many faces of instructional design: Toward a framework for the design of powerful learning environments.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? Cognition and Instruction, 21(3), 209–249.

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes. Harvard university press.

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

Weinberger, A., Fischer, F., & Mandl, H. (2002). Fostering computer supported collaborative learning with cooperation scripts and scaffolds.

Wertsch, J. V. (1986). Culture, communication, and cognition: Vygotskian perspectives. CUP Archive.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. Cognition and Instruction, 16(1), 3–118.

Wick, M. R., & Slagle, J. (1989). The partitioned support network for expert system justification. IEEE Transactions on Systems, Man, and Cybernetics, 19(3), 528–535. https://doi.org/10.1109/21.31059

Wick, M. R., & Thompson, W. B. (1992). Reconstructive expert system explanation. Artificial Intelligence, 54(1–2), 33–70.

Zagalsky, A., Teshima, C. G., German, D. M., Storey, M.-A., & Poo-Caamaño, G. (2016). How the R community creates and curates knowledge: A comparative study of stack overflow and mailing lists. Proceedings of the 13th International Conference on Mining Software Repositories, 441–451. https://doi.org/10.1145/2901739.2901772

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.