Human-AI Collaboration for UX Evaluation: Effects of Explanations and Synchronization

MINGMING FAN*, The Hong Kong University of Science and Technology, China
XIANYOU YANG†, Rochester Institute of Technology, USA
TSZ TUNG YU†, University of Waterloo, Canada
Q. VERA LIAO, Microsoft Research, Canada
JIAN ZHAO, University of Waterloo, Canada

Analyzing usability test videos is arduous. Although recent research showed the promise of AI in assisting with such tasks, it remains largely unknown how AI should be designed to facilitate effective collaboration between user experience (UX) evaluators and AI. Inspired by the concepts of *agency* and *work context* in human and AI collaboration literature, we studied two corresponding design factors for AI-assisted UX evaluation: *explanations* and *synchronization*. Explanations allow AI to further inform humans how it identifies UX problems from a usability test session; synchronization refers to the two ways humans and AI collaborate: *synchronously* and *asynchronously*. We iteratively designed a tool—AI Assistant—with four versions of UIs corresponding to the two levels of explanations (with/without) and synchronization (sync/async). By adopting a hybrid wizard-of-oz approach to simulating an AI with reasonable performance, we conducted a mixed-method study with 24 UX evaluators identifying UX problems from usability test videos using AI Assistant. Our quantitative and qualitative results show that AI with explanations, regardless of being presented synchronously or asynchronously, provided better support for UX evaluators' analysis and was perceived more positively; when without explanations, synchronous AI better improved UX evaluators' performance and engagement compared to the asynchronous AI. Lastly, we present the design implications for AI-assisted UX evaluation and facilitating more effective human-AI collaboration.

CCS Concepts: • **Human-centered computing** → **Usability testing**; **Empirical studies in HCI**.

Additional Key Words and Phrases: human-AI collaboration, user experience (UX), AI-assisted UX evaluation, explainable AI, intelligent user interface design, synchronization, explanations, think-aloud usability test

ACM Reference Format:
Mingming Fan, Xianyou Yang, Tsz Tung Yu, Q. Vera Liao, and Jian Zhao. 2022. Human-AI Collaboration for UX Evaluation: Effects of Explanations and Synchronization. *Proc. ACM Hum.-Comput. Interact.* 00, 00, Article 00 (2022), 32 pages. https://doi.org/10.1145/1122445.1122456

*Corresponding author
†equal contribution

Authors' addresses: Mingming Fan, mingmingfan@ust.hk, The Hong Kong University of Science and Technology, China; Xianyou Yang, xy1258@rit.edu, Rochester Institute of Technology, USA; Tsz Tung Yu, tt3yu@uwaterloo.ca, University of Waterloo, Canada; Q. Vera Liao, veraliao@microsoft.com, Microsoft Research, Montreal, Canada; Jian Zhao, jianzhao@uwaterloo.ca, University of Waterloo, Canada.

2573-0142/2022/00-ART00 $15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Usability testing has been widely adopted in both industry and academia to identify user experience (UX) problems when developing interfaces and systems [25, 58]. With the convenience and easy access to diverse users, remote usability testing, by conducting and recording the tests through video conferencing tools and analyzing the recorded sessions afterwards, has become increasingly popular [3, 16, 25]. However, analyzing recorded usability test sessions is labor-intensive and time-consuming, as UX evaluators have to listen to what users verbalize while watching their actions to catch sporadic problems in an often lengthy video. In practice, UX evaluators often have to complete the analysis in a short period [27] and tend to perform quick rather than thorough analysis. What's more, UX evaluators who analyze the same usability test sessions have been found to identify substantially different sets of usability problems, which is known as the "evaluator effect" [38]. One important way to cope with the "evaluator effect" is to involve multiple evaluators to analyze the same test session. However, due to practical constraints (e.g., limited company resources), few evaluators (23%) had a chance to gain a different perspective from other *human* evaluators on the same usability test session [27]. One possible solution, in light of recent advancement in AI, is to investigate whether an AI agent could be designed to analyze the same usability test session and provide a different perspective to the human evaluator.

Toward this goal, recently researchers began to explore human-AI collaborative approaches, in which AI plays the role of another evaluator and suggests potential usability problems to the UX evaluator. For example, VisTA [26] is an analytical tool that visualizes AI predicted problems before a UX evaluator starts their analysis; thus, the collaboration between the AI and the evaluator is *asynchronous*. Such asynchronous human-AI collaboration was also explored in other contexts [36, 75, 76, 80, 88, 90, 92, 97]. In human-human collaboration, synchronization, whether the collaboration happens *asynchronously* and *synchronously*, is considered a key dimension that impacts the user requirements and design of CSCW systems, as seen in the *Time* dimension of Johansen's Time-Space Matrix [45]. To our knowledge, *synchronous* collaboration between humans and AI is relatively understudied in knowledge-rich domains; What's more, no research has compared asynchronous and synchronous collaboration in the context of usability evaluation. We took a first step to investigate how **synchronization** between the AI and UX evaluators might affect their collaboration experiences.

Furthermore, when analyzing a usability test video, UX evaluators care about not only *where* in the video the AI believes the user encounters a problem but also *why* it believes so–the **explanations** of AI's judgment. Such an explanation could allow evaluators to better assess whether to incorporate or reject AI's judgement [26]. On the other hand, recent work on "agency divide" between human and AI [49, 50] suggests that providing explanations could affect human's perceived agency. Specifically, humans perceive full agency of their work when there is no AI assistance, and decreasing agency as the amount of information and assistance provided by the AI increases, including explanations. A decreased human agency could negatively affect the engagement and outcome of usability video analysis. Given these potential effects, we empirically investigated how accessing AI's explanation, and how explanation and synchronization working together, affect human-AI collaboration in the context of usability test video analysis.

Towards these study goals, we iteratively designed *AI Assistant*, a tool assisting UX evaluators with analyzing usability test videos. As it is still challenging to develop an AI that could detect UX problems of any given usability test video with consistent accuracy, we adopted a wizard-of-oz (WoZ) approach to simulating an AI that could detect UX problems with reasonable precision and recall. We also designed WoZ explanations based on both what matters in analyzing usability test video [26] and what is potentially feasible with state-of-the-art explainable AI (XAI) techniques.

We designed AI Assistant with four different versions of user interfaces (UIs), which present the AI-suggested usability problems *synchronously* or *asynchronously* and *with* or *without* explanations. We conducted a mixed-methods study with 24 UX evaluators, in which they analyzed two usability test videos with AI Assistant and were interviewed afterwards regarding their experiences, perceptions, and preferences of AI Assistant.

Our results showed that both synchronization and explanations positively affected human-AI collaboration. Compared to the asynchronous AI without explanations, AI with explanations, regardless of being presented asynchronously or synchronously, helped to improve UX evaluators' performance of (e.g., the number of identified UX problems) and engagement (e.g., time spent) in their analysis, and their perception of AI Assistant (e.g., understanding of AI). When without explanations, the synchronous AI helped to improve UX evaluators' performance and engagement more than the asynchronous AI. Our qualitative results from semi-structured interviews provide further insights into the effects of the two factors. Based on the findings, we discuss the design implications of explanations and synchronization for UX evaluation and effective human-AI collaboration.

2 BACKGROUND AND RELATED WORK

2.1 AI for Usability Problems Detection

Analyzing usability test video to identify problems that participants encountered is a common task for UX evaluators. They need to attend to multiple behavioral signals of the participant from both visual and audio channels of the test video. More importantly, they need to leverage their domain expertise to determine whether there is indeed a usability problem or whether it is just typical efforts (e.g., trial-and-error) that the participant had to make when using the product. To help UX evaluators better analyze usability test videos, recent research began to develop AI methods to predict the overall UX of interfaces [65] or detect specific usability problems [32, 35, 44, 67]. For example, Grigera et al. proposed a rule-based classifier to detect a predefined set of usability smells—the hints of bad designs that could cause usability problems—by analyzing users' interaction logs [32]. Similar rule-based classification approaches have been used to detect usability smells in mobile websites [67] and in virtual reality applications [35]. Jeong et al. proposed a graph-based AI method to model and measure the similarity of users' interactions with a mobile application to detect potential usability problems [44]. Although such automatic methods show the promise of detecting simple usability problems for specific user interfaces based on structured data (e.g., interaction logs) and empirical rules, detecting usability problems directly from a usability test video, which is a primary task of UX practitioners, is still challenging to be fully automated as it requires an understanding of the functions and designs of the test product, the test tasks, and the test subject's behaviors.

To address the limitation of fully automated methods, researchers began to investigate human-AI collaboration tools to support UX evaluators rather than replacing them. One representative work is VisTA, a visual analytical tool that integrates visualization and machine learning (ML) to detect and highlight segments of a usability test video containing potential usability problems [26]. Such human-AI collaboration was shown to help UX evaluators identify more problems than they worked alone [26]. As a first step to experiment human-AI collaboration for usability test video analysis, VisTA was limited in two ways. First, the collaboration between the AI and the UX evaluator was *asynchronous*. The evaluator was shown with the AI's predictions before they started their analysis, which might have affected the evaluator's independent analysis. Alternatively, the AI's predictions could be shown *synchronously*, creating a perception that the evaluator and the AI are analyzing a usability test video simultaneously, which might allow the evaluator to perform more independent analysis while still being able to access the second perspective from the AI.

The second limitation, based on their study participants' feedback, was that VisTA remained an "opaque box" and did not allow an understanding of how the AI worked. Although VisTA visualizes the input features to show *what* went into the AI's analysis, it was deemed insufficient without having access to the meaning of these input features or the rationales on which the judgments were based. Thus, it is necessary to explore explanations that could answer the *why* question in identifying usability problems in a video, such as what design principles or usability heuristics were violated [26].

Our work seeks to understand how synchronization (i.e., synchronous and asynchronous collaboration) and explanations (i.e., with and without) would affect UX evaluators' collaboration with the AI in the context of usability test video analysis. To overcome the technical limitations of AI in detecting UX problems, we opt for a hybrid Wizard-of-Oz (WoZ) approach in order to focus on evaluating different ways to present the AI's suggestions with a controlled experiment. Our approach simulates a reliably performing AI system but is grounded in the technical feasibility of ML technologies. Specifically, we assume the system works with a supervised ML classifier, which takes a segment of a usability test video as input and predicts whether the test subject encounters a problem in that video segment. The design of our WOZ AI's functionalities, including the AI explanations, will be discussed in Section 3.3.2.

2.2 Human-AI Collaboration

The term "human-AI collaboration" has emerged in recent research studying the usage of AI systems [5, 14, 87]. There is a shift from an "automation" perspective of AI to recognizing that AI systems should be used to support, instead of replacing, the decisions and tasks of domain workers. For example, an "algorithm-in-the-loop" process [31] is adopted in many high-stakes decision-making contexts, where ML models are used to inform people its assessment or to alert cases falling in a targeted category. Ultimately, people have to decide whether to accept the AI's recommendations or discount them [95]. Following this collaborative perspective, researchers studied how people perceive algorithmic decision assistance [10, 52], how they use AI systems for decision-making [13, 14], often in unexpected or suboptimal ways [18, 93], and how to improve the human-AI joint decision-making outcomes [8, 89, 95].

While research has only begun to define and study different types of human-AI partnership, it would be advisable to draw on theories and insights from human-human collaboration research. CSCW research has identified a number of core dimensions to characterize different collaborative tasks. One is the division of work and labor between parties [82]. Recent work by Lai et al. [49, 50] proposed a notion of "divide of agency" between human and AI with a spectrum from full human agency to full automation. In particular, AI that provides *explanations* for its predictions is considered to allow lower human agency than one that offers predictions only. On the other hand, explanations not only could affect the perception of AI's competence [73, 74], but also provide additional assistance for decision-making, for example, by highlighting important factors in the decision or illuminating useful rationales [50, 54]. However, such improvements from explanations were typically observed with the AI that outperformed the human. Bansal et al. recently examined how the explanations of the AI with human-comparable performance affect human-AI collaboration [9]. Our work extends theirs in two aspects. First, in Bansal et al.'s studies, AI's explanations were presented to human participants upfront before they started their own analysis. Such asynchronous collaboration, as they acknowledged, made it almost impossible for the participants to reason independently. In contrast, our work studies the potential effects of synchronization between humans and the AI. Second, Bansal's studies used tasks amenable to crowdsourcing, and the findings might not be generalizable to experts in high-stake scenarios. In contrast, our tasks require domain expertise in user experience (UX) research. Thus, the type of human-AI collaboration in

our work is between AI and domain experts, and the findings are complementary to those of Bansal et al.'s.

The WoZ explanations of AI assistant in this research is informed by the methods developed in the field of Explainable AI (XAI); A full review of prior work on this topic is beyond the scope of this paper but can be found in many recent survey papers [1, 6, 15, 30, 34]. Explanations could be generally categorized into *global* and *local* explanations. While global explanations describe how a model makes a decision in general, such as how it weighs different features and what rules it follows, local explanations focus on justifying a decision made for a particular instance, for example, by highlighting important features of the instance that contributed to the AI's decision [6, 34]. The explanations presented in our AI assistant is an example of local explanations, which highlights exceptional features in the usability test video that AI assistant considers as the indicators of a usability problem. To make AI explanations more accessible to laypeople, an emerging area of XAI work explored *explanation generation* using domain-specific semantics [39, 46] or rationales [21], enabled by additional human supervision or training data. For example, trained, on human explanation data, rationale generation [21] translates AI's internal representation into natural-language rationales.

Another pivotal dimension to characterize collaborative tasks is the *Time* dimension of the well-known Time-Space matrix of human-human collaboration [45]. The Matrix characterizes collaborative tasks and computing tools to support workers by a *Time* dimension—whether individuals collaborate *synchronously* or *asynchronously*, and a *Space* dimension—whether collaboration is co-located or geographically distributed. While the Space dimension is irrelevant to human-AI collaboration, the Time dimension is pertinent. However, to our knowledge, little is understood about how human-AI collaboration is impacted by the Time dimension. However, these factors could produce distinct interaction experiences that may impact how one perceives and interacts with AI, and ultimately the collaborative outcomes. While our current research is carried out in the context of UX evaluation, *synchronization* is broadly applicable to AI-assisted decision tasks.

2.3 Collaborative Analysis Tools

Previous research has investigated ways to support collaboration between humans, including design principles (e.g., time and space model [42, 45] and information scent [70]), collaborative data analysis and sensemaking [19, 42, 68] as well as collaborative infrastructures and tools. Examples of collaborative infrastructures include ManyEyes [86] for sharing and commenting on data charts, Polychrome [7] for collaborative web visualizations, and BEMViewer [59] for supporting branch-explore-merge protocol in data exploration. Collaborative tools support human-human collaborations for synchronous and asynchronous tasks. Synchronous collaborative tools focus primarily on increasing coworkers' awareness [96], building common ground [57], and sharing/constraining personal workspaces [43, 84]. Asynchronous collaborative tools often address the design challenges in facilitating communication [36, 88], supporting handoff [97], reasoning actions [90], and viewing analysis histories [75, 76, 92].

Previous research has primarily focused on supporting collaboration between humans in non-UX domains. In this research, we draw inspirations from this body of literature to design an AI-assisted tool to support synchronous and asynchronous collaboration between AI and UX evaluators when they analyze usability test videos. This tool is then used as the vehicle to study the effect of the two factors—synchronization and explanations—on UX evaluators' workflow.

3 METHOD

The goal of this research is to understand the effect of synchronization and explanations on human-AI collaboration for UX research in the context of analyzing usability test videos. The findings would provide design implications for the two factors in human-AI collaboration.

3.1 Research Questions

To achieve the above overarching goal, we seek to answer three research questions (RQs):

RQ1: How would the *explanations* of AI affect UX evaluators' performance and perception in analyzing usability test videos?

RQ2: How would the *synchronization* of AI affect UX evaluators' performance and perception in analyzing usability test videos?

RQ3: What are UX evaluators' preferences that can inform the design of AI tools assisting UX evaluation?

3.2 *AI Assistant* for Usability Test Video Analysis

Usability test video analysis requires UX evaluators to review the video recordings of test sessions to identify UX problems. We designed a tool—*AI Assistant*—to support the task by suggesting potential UX problems in the video. It supports watching a usability test video, identifying and annotating UX problems, and viewing AI assistance in one place, with the following UI components:

Video Player. AI Assistant includes a basic video player, with which UX evaluators can play, pause, rewind, and jump forward a video (Figure 1a).

Annotation Panel. After identifying a usability problem from watching the test video, UX evaluators often need to describe the problem for reporting purpose or future reference. AI Assistant includes an annotation function, which shows the current timestamp of the video and contains a text field for writing problem description. Since the annotation function would be used frequently, we position it side-by-side with the video (Figure 1b).

Problem Table. To allow UX evaluators to view the problems that they have identified, AI Assistant shows all the problems and the corresponding descriptions in a table (Figure 1c).

AI-Suggested Problems Timeline. One key function of AI Assistant is introducing AI to assist UX evaluators with their analysis. Specifically, the AI analyzes the video and infers whether and when (i.e., a time duration) the user in the video is encountering a usability problem. The AI-suggested UX problems are visualized as a horizontal timeline chart, with the value `1` representing a problem being encountered and the value `0` representing no problem being encountered (Figure 1d). The timeline is positioned below the video to facilitate UX evaluators to glance at it during their analysis. Detailed design of the timeline will be introduced in Section 3.4.

AI Explanations Panel. For the versions of AI Assistant with explanations, the tool features a separate panel to show the explanations of why the AI infers there is a UX problem in a given time duration. The explanations is shown below the timeline (Figure 1e). The content and design of the explanations will be described in Sections 3.3 and 3.4.

As discussed, given the technical challenge to train an ML model that maintains consistently satisfying performance for different usability test videos, we utilized a *wizard-of-oz (WoZ)* approach to generating AI-suggested problems, serving as the back-end AI of the tool. An WoZ approach also allowed us to control the content of explanations to make them consumable for UX evaluators.

3.3 Task Videos and WoZ AI

3.3.1 Task Videos. We needed usability test videos for UX evaluators to analyze in our experimental study. To do so, we selected two recorded videos with audio from a dataset curated in our previous

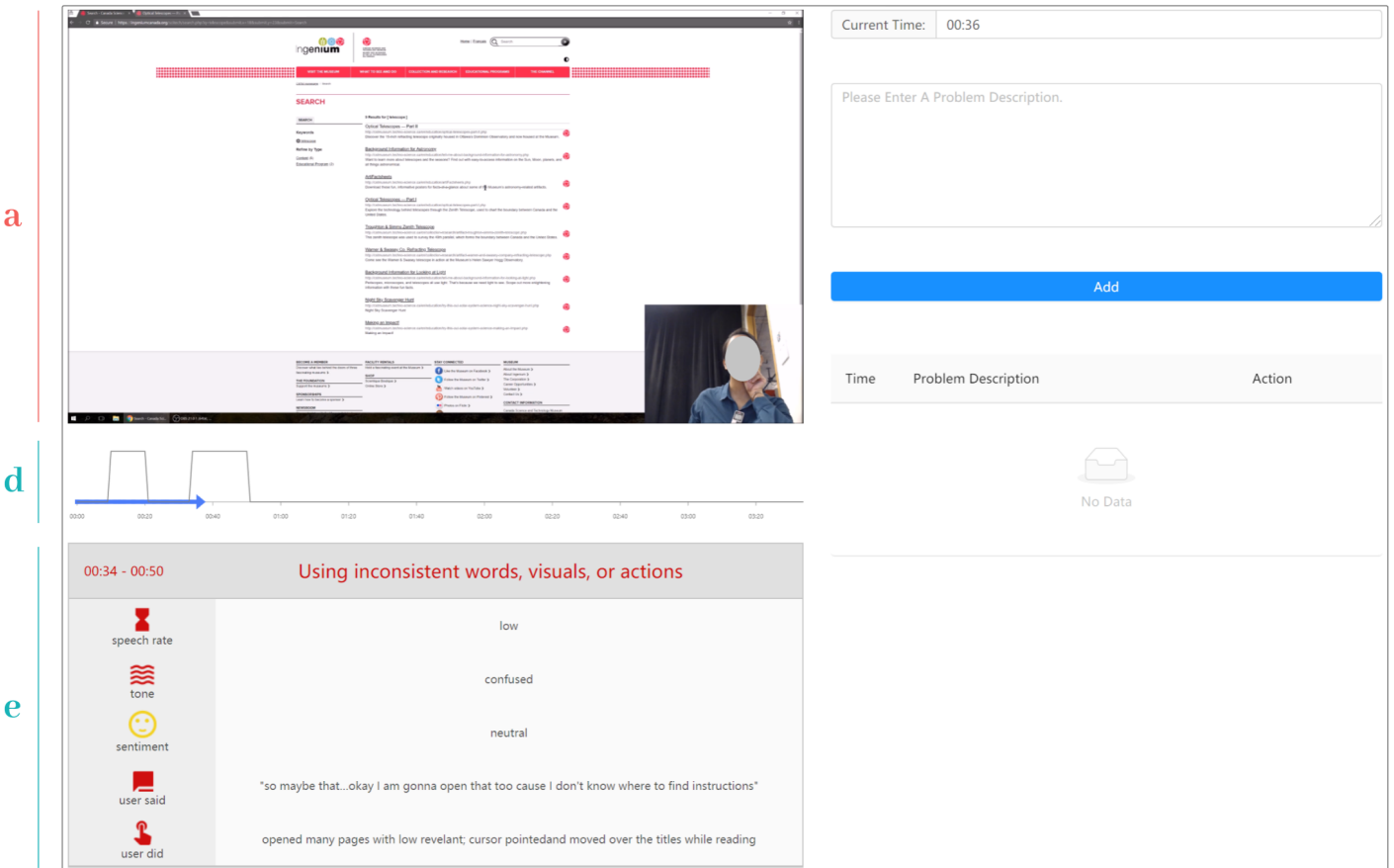

Fig. 1. The user interface of AI Assistant that presents the AI-suggested problems *synchronously with explanations*: (a) video player; (b) annotation panel; (c) identified problem table; (d) timeline of the AI-suggested problems; (e) explanations of the AI-suggested problems.

studies [24]. To create the dataset, we recruited eight native-English speakers (four females and four males, aged 19-26) to participate in usability test sessions, in which they performed tasks on two websites and two physical devices while thinking aloud. All test sessions were video and audio recorded. One of the selected videos was about a participant browsing a website of a national science museum. The task was to find a photo of the instructions to operate an early telescope. The other selected video was about a participant using a multi-function coffee machine, and the task was to program the machine to make two cups of strong-flavored drip coffee at seven thirty in the morning. The website video lasted 214 seconds, and the coffee machine video lasted 682 seconds. We intentionally selected videos with different interfaces (virtual and physical) as well as different lengths (short and long) to cover a range of scenarios. Each video was seen by participants as one evaluation task, presented in the video player panel in Figure 1.

3.3.2 Wizard-of-Oz (WoZ) AI. We introduce the details of our hybrid WoZ AI and discuss potential technical feasibility below. We will reflect on the limitations of WoZ and future work in Sec. 6.4.

There are two pieces of information that the WoZ AI would offer to UX evaluators: *AI-suggested problems* and *AI explanations* to these problems. To generate AI-suggested problems, we first needed to find the *ground-truth problems* in the usability test videos.

Ground-truth Problems. To identify the ground-truth UX problems that users encountered in the two selected usability test videos, nine UX evaluators were recruited to independently review the videos, identify segments in the videos where users encountered problems, and write descriptions for the problems. Next, two other UX researchers reviewed the identified problems and their descriptions, discussed and consolidated a final list of ground-truth problems and descriptions. In total, there were 20 UX problems in the coffee machine video, and 8 in the website video.

Supervised ML Predicting Usability Problems in Video Segments Our WoZ system was assumed to work with a supervised ML classifier trained on video data with labels of whether a segment contains a usability problem. Typically, the input features of a video-based supervised ML classifier include acoustic features (e.g., pitch) from the audio track, textual features from the transcription, and visual features from the video. To allow more human-consumable explanations, we also assume that more high-level semantic features, such as user actions and sentiments in the texts, can be obtained by either automatic recognition techniques or additional supervision

It is common sense that AI would hardly be perfect, especially in knowledge-rich domains, such as usability testing. Thus, it would be inappropriate to have the WoZ AI suggest all the ground-truth problems. To make the WoZ AI more realistic, we randomly added 5 (18%) false problems (false positives) and removed 4 (14%) true problems (false negatives) in the two selected videos in total.

AI Explanations. Our design of AI explanations is based on the potential technical feasibility of explainable AI, and informed by how usability problems are identified and explained in UX practices [25, 27, 58]. Two forms of explanations are often generated by XAI techniques to explain an AI's decision for a particular instance [1, 6, 15, 30, 34]: *rule-based*—the decision rules that this instance violates or complies with; *feature-based*—the features of this instance that are strong indicators for the decision. Correspondingly, we explain an AI-suggested problem in a usability test video segment by showing the *UX design heuristics* that the video segment suggests the test product may violate, and the *behavioral features* of the user that are indicative of the usability problem. These two types of information about AI's predictions were also desired by UX evaluators when they worked with an AI agent to analyze usability test videos [26, 80]

UX Design Heuristics. UX literature has offered heuristics and principles for designing good user interfaces and identifying bad ones, such as Nielsen's heuristics [63], Norman's design principles [64], and Gerhardt-Powals's cognitive engineering principles [28]. As these principles and heuristics are largely overlapped, we employed Nielsen's heuristics as they are commonly used

in industry. The original Nielsen's heuristics are abstract and require explanations to understand properly. As a result, we revamped the heuristics to make them more self-explainable while still keeping them short by referring to the explanations given by the Nielsen and Norman Group [33]. For example, the fourth heuristic is "consistency and standards." The explanations of the violation of this heuristic is "Using inconsistent words, visuals, or actions." To identify the violated heuristics for AI-suggested problems, we consulted the problem descriptions consolidated for the ground-truth problems and mapped them to the closest usability heuristics.

There are several potential ways to implement such explanations. The heuristics themselves can be included as high-level features in the usability problem detection model and then directly provided as explanations. We can train additional supervised models to recognize these heuristics features. Recent research also began to investigate ways to automatically detect the violations of usability heuristics. For example, Ponce et al. [71] proposed a convolution neural network model to detect the violation of three of the Nielsen's heuristics [63]. Alternatively, one can leverage rationale or explanation generation techniques trained on human explanation data [21, 39] to generate these design heuristics.

Behavioral Features. The AI explanations also include the following behavioral features as indicators of the UX problems: *speech rate*, *tone (i.e., pitch)*, *speech sentiment*, *speech content*, *what the user did in the video*. These features are chosen for two reasons. First, previous research suggests that UX evaluators should pay attention to what and how the user verbalizes and what the user does on the test product to identify UX problems [25, 58]. For example, when encountering problems, users may slow down their speech, raise their tones, or use words with negative sentiments more often [22, 24, 48, 69, 81]. Second, it is possible for current or near-future sensing and AI technologies to capture these behavioral features from usability test videos. For example, there is a rich body of literature and commercially available solutions for *emotion* (e.g., confusion) and *sentiment* analysis (e.g., iMotions [41]). Computer vision techniques, especially work on natural language video description [4, 85], can be used to identify and describe *what the user did* in video segments.

For WoZ AI, we adopted a hybrid approach to generate the behavioral features to explain the suggestion of a UX problem. The feature categories of *speech rate*, *tone*, and *sentiment* were automatically extracted with the following algorithm. The audio track of each usability test video was divided into small segments based on silences in the user's think-aloud verbalizations and then transcribed into texts. For each segment, we calculated the speech rate by dividing the number of words spoken in the segment by its duration. We calculated the tone (i.e., pitch) of the user by computing the fundamental frequency F0 (Hz) at the sampling rate of 100 Hz using the praatUtil library [37]. For the sentiment, we analyzed the transcript using the VADER library [40] and then discretized it into three levels: negative, neutral, and positive. For the category of *what user did*, we manually reviewed the videos and created a description.

3.4 Design for Synchronization and Explanations

As we were interested in studying the synchronization and explanations factors of AI, we designed AI Assistant with four alternative UIs to assist UX evaluators through an iterative design process. The four versions include: (a) synchronously presenting the AI-suggested problems *with* explanations (*w/, sync*), (b) synchronously presenting the AI-suggested problems *without* explanations (*w/o, sync*), (c) asynchronously presenting the AI-suggested problems with explanations (*w/, async*), and (d) asynchronously presenting the AI-suggested problems without explanations (*w/o, async*). Figure 1 shows the entire interface of the *(w/, sync)* version of the AI Assistant. Figure 2 shows the differences (in the bottom left part of Figure 1de) among the four versions. The (*w/o, async*) condition (Figure 2d) is the *baseline* since it is the simplest way of presenting UX problems among the four conditions.

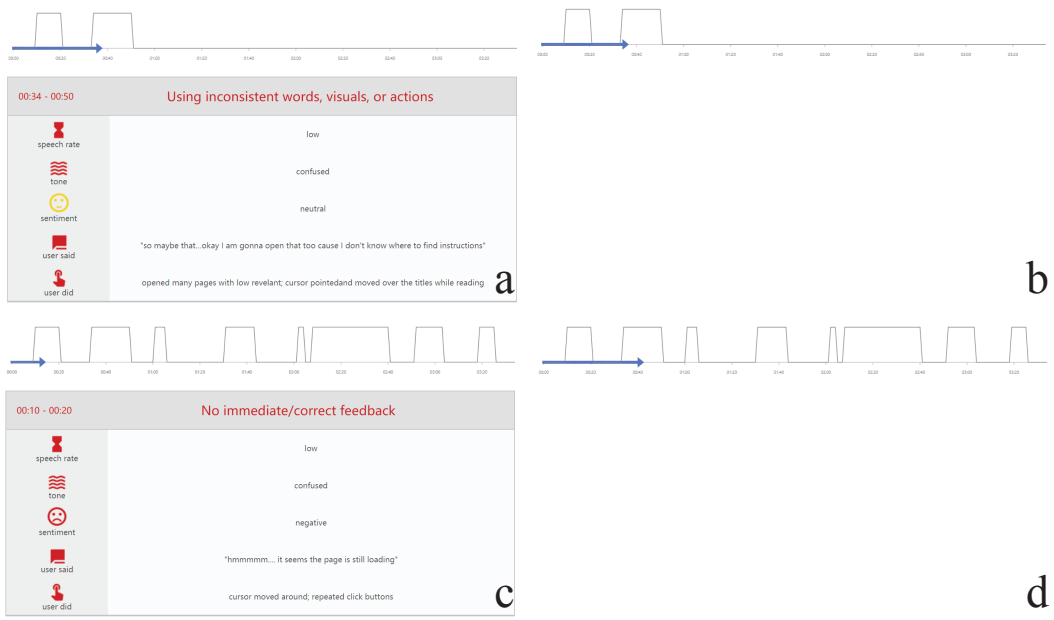

Fig. 2. Comparison of the four different versions of AI Assistant (the bottom-left part of the user interface, as shown in Figure 1de): (a) synchronously presenting the AI-suggested problems *with* explanations (*w/, sync*), (b) synchronously presenting the AI-suggested problems *without* explanations (*w/o, sync*), (c) asynchronously presenting the AI-suggested problems with explanations (*w/, async*), and (d) asynchronously presenting the AI-suggested problems without explanations (*w/o, async*), which is the baseline.

3.4.1 Design for Synchronization. The difference between the *async* and the *sync* versions of AI Assistant lies in how the AI-suggested problems are revealed to UX evaluators.

In the *async* way of presenting the AI, the AI-suggested problems are revealed to UX evaluators all at once at the beginning of their analysis and are always available during their analysis (Figure 2cd). This design simulates an *asynchronous* workflow in which the AI completed its analysis before the UX evaluator. A similar timeline was used in the literature [26].

In the *sync* way of presenting the AI, the AI-suggested problems are revealed to UX evaluators progressively as they play the video. This design simulates a *synchronous* workflow in which the AI and the UX evaluator work through the video together at the same time (Figure 2ab). UX evaluators are able to see the AI's suggestions up to the point where the video has been playing to, and cannot see the AI's suggestions for the rest of the video. In cases where UX evaluators skip a portion of the video and start to play the video from a later timestamp, the AI's suggestions for the skipped portion are not revealed to them. But the AI's suggestions from that future timestamp onward will be progressively revealed to the UX evaluators as they continue playing the video. We made this design choice to avoid UX evaluators from gaming the system by jumping to a very late timestamp of the video with the hope of revealing all the AI's suggestions.

3.4.2 Design for Explanations. For the *w/* problem explanations versions of AI Assistant, the AI explanations Panel (Figure 2ac) contains two types of information: the UX design heuristics that are violated and the behavioral features that suggest UX problems, as described in Section 3.3.2. The key design consideration for presenting the explanations content is to ensure it is *light-weight* and *scannable*. UX evaluators need to attend to different types of information—the video to be

analyzed and the AI-suggested problems—while writing their own problem descriptions using the annotation function. Thus, the visual representation of the explanations should minimize the attention needed and be scannable without extensive reading.

Following the design consideration, we minimized the amount of text by using visual icons that conveyed information effectively and organized the features in a consistent and scannable format.

3.5 Implementation

We implemented the four versions of AI Assistant as a web application, which would allow for recruiting a broad set of UX evaluators to participate in the study. We deployed the tool on Google Firebase and the videos on Amazon Web service (AWS) to allow UX evaluators to access the tool remotely and robustly.

The tool includes a back-end NodeJS server and a front-end UI that is built on top of React and Mobx libraries. The NodeJS server provides an application programming interface (API) to serve the links of the usability test videos from Amazon S3 to our front-end UI seamlessly. The AI-suggested problems and the corresponding explanations for each usability test video are stored into a JSON file, which is visualized on the front-end UI in real-time depending on the experimental condition. The front-end UI has been revamped by utilizing React Router into the application, which allows us to control the experimental conditions that UX evaluators have access to by sharing them with the corresponding URLs. By using Mobx, the activity history of the video player (e.g., how the video is played, paused, rewind) and the usability problems identified by UX evaluators are stored in the state during the user study and saved into the JSON format log data, which will be analyzed as part of the quantitative data of the study.

4 USER STUDY

To answer the RQs, we used a mixed-method, which included a controlled user study, questionnaires, and interviews. In the controlled study, UX evaluators analyzed two usability test videos with AI Assistant and answered questionnaires regarding their use of AI Assistant. Post-study interviews offered more insights into their preferences of AI Assistant.

4.1 Participants

We recruited participants from industry and local universities via electronic flyers with two inclusion criteria: having at least one year's experience in UX and having experience of analyzing think-aloud usability test sessions. We first conducted a pilot study with two participants, who had 1-2 years of UX/HCI experience. The two pilot study sessions helped us to adjust the whole study procedure and duration, ensuring that the study design was appropriate.

We then recruited 24 participants as UX evaluators for the actual user study. Among the 24 participants, sixteen were females and eight were males; thirteen were at the age of 25-34 and eleven at the age of 18-24. In terms of educational backgrounds, fifteen had Bachelor's degrees, four had Master's degrees, two had Doctoral degrees, and three had some college degrees. Regarding the UX/HCI work experience, fourteen participants had 1-2 years of experience, six had 2-3 years of experience, and four had more than 3 years of experience.

4.2 Experimental Design

We employed a 2-by-2 mixed design, with the *explanations* as the between-subjects factor and the *synchronization* as the within-subject factor. This means each participant was randomly assigned to use AI Assistant either *w/* explanations or *w/o* and to complete two usability video analysis tasks, one of which was with the *sync* AI Assistant and the other was with the *async*. We counter-balanced the order of the *sync* and *async* AI Assistants and randomized the order of the two videos.

Table 1. Participant information and the study design. For trust in AI, higher score indicates more trust; for experimental conditions, video-1 is the physical device video and video-2 is the digital website video.

ID	Knowledge about Nielsen's UX design Heuristics [63]	Knowledge about AI	Trust in AI	Experimental conditions
P1	Some	A little	4	(w/,async, video-1), (w/, sync, video-2)
P2	Some	A little	4	(w/o, sync, video-1), (w/o, async, video-2)
P3	A lot	A little	4	(w/o, async, video-1), (w/o, sync, video-2)
P4	Some	A little	5	(w/, sync, video-2), (w/, async, video-1)
P5	A lot	A little	5	(w/, async, video-2), (w/, sync, video-1)
P6	A lot	A little	4	(w/o, sync, video-2), (w/o, async, video-1)
P7	Some	No knowledge	4	(w/o, async, video-2), (w/o, sync, video-1)
P8	Some	A little	3	(w/, sync, video-1), (w/, async, video-2)
P9	A lot	A little	4	(w/, async, video-1), (w/, sync, video-2)
P10	Some	A little	3	(w/o, sync, video-1), (w/o, async, video-2)
P11	Some	No knowledge	2	(w/o, async, video-1), (w/o, sync, video-2)
P12	A lot	Some	4	(w/, sync, video-2), (w/, async, video-1)
P13	A lot	A little	3	(w/, async, video-2), (w/, sync, video-1)
P14	Some	Some	3	(w/o, sync, video-2), (w/o, async, video-1)
P15	Some	A little	4	(w/o, async, video-2), (w/o, sync, video-1)
P16	Some	Some	4	(w/, sync, video-1), (w/, async, video-2)
P17	A lot	Some	3	(w/, sync, video-1), (w/, async, video-2)
P18	Some	A little	4	(w/, async, video-1), (w/, sync, video-2)
P19	A lot	A little	4	(w/o, sync, video-1), (w/o, async, video-2)
P20	A lot	A little	3	(w/o, async, video-1), (w/o, sync, video-2)
P21	A little	Some	2	(w/, sync, video-2), (w/, async, video-1)
P22	Some	A little	3	(w/, async, video-2), (w/, sync, video-1)
P23	A lot	A little	3	(w/o, sync, video-2), (w/o, async, video-1)
P24	A lot	A little	4	(w/o, async, video-2), (w/o, sync, video-1)

We chose this mixed design for the following reasons. First, there is a potential learning effect for explanations. After participants see the explanations for the first test video, they might be primed to use similar information when analyzing the second video even if they are not given explanations. Thus, we set the explanations as the between-subjects factor. Second, there is no foreseen learning effect for synchronization so it can be set as the within-subject factor. Doing so also allows each participant to compare their experiences with *sync* and *async* human-AI collaboration.

4.3 Procedure

The study was divided into the following phases: *set-up*, *pre-test questionnaire*, *training session*, *two formal-task sessions* (each including one formal task and one post-task questionnaire), and *post-test questionnaire and interview*. All studies, including the pilot studies, were conducted online. The study took 75 minutes on average to finish, and each participant was compensated with $15.

Set-up. Participants were asked to enter our ZOOM room, check their microphone, speaker and camera, and to share their screen with the moderator.

Pre-test Questionnaire. The moderator sent participants the URL to the online questionnaire via ZOOM chat. In the pre-test questionnaire, participants were asked to fill in their basic demographic questions (e.g., age, gender, education backgrounds), years of UX/HCI work experience, and their experience of conducting usability testing sessions. Since they would analyze the usability test videos with the AI, they were also asked about their knowledge of AI and ML (four levels from no knowledge to a lot of knowledge). Moreover, they were also asked to rate their general trust in AI and ML on a 5-point Likert scale.

Training Session. We asked participants to learn and practice using the AI assistant to analyze a usability test video different from the ones used in the formal task phase. We first introduced to participants how to play the video and write problem descriptions. We then explained the visualization of the AI's inferences. For the participants in *(w/, sync)* and *(w/, async)* conditions that showed the AI's explanations, we explained to them that the explanations consisted of 1) the UX design heuristics that were violated; and 2) the input features that the AI considered when making its predictions. We kept the introduction consistent for all participants in these conditions. We informed participants that they could decide whether to use the AI or not. We then showed them with an example of how to find and record a usability problem. Next, we asked participants to practice using the tool to find and record one usability problem on the practicing video.

Formal Task Sessions. Participants used two versions of AI Assistant to review the two usability test videos (described in Section 3.3) to identify problems and write problem descriptions. As discussed in the previous section, one task was with the *sync* AI Assistant and the other with the *async* AI Assistant, and the order of the two versions of AI Assistant was counter-balanced. For each usability test video, the moderator suggested (but not enforced) participants spend no more than about two times of the video length to keep the study on time. After completing each task, the participants were asked to complete a post-task questionnaire.

Post-test Questionnaire and Interview. After finishing reviewing each of the two usability test videos, the participants were asked to complete a post-task questionnaire reporting their satisfaction, understanding, and trust in AI Assistant (details in Section 4.4). After completing both tasks, we conducted semi-structured interviews for 10-20 minutes regarding participants' experiences (e.g., how did you use AI Assistant? what did you like and dislike about AI Assistant? why?) and preferences (e.g., which version of AI Assistant did you prefer more? what other features of AI Assistant would you want? why?) of using AI Assistant, as well as other customized questions regarding the moderator's observations of their behaviors in the study.

4.4 Measurements

From the experiment, we captured three types of measurements, including task performance metrics related to identified UX problems, behavioral metrics reflecting interaction patterns, and subjective perception of AI Assistant measured by survey responses.

Task Performance. This type of measures focused on the number of UX problems found by each participant. Given the ground truth of UX problems in the two usability tasks and the problems identified by a participant, we calculated the *precision*—the percentage of correct problems among all problems identified by the participant, and *recall*—the percentage of correct problems identified by the participant among all correct problems existed in the ground truth. To further verify the effect of AI Assistant on participants' task performance, we also analyzed the *overlap* between a participant's identified problems and AI Assistant's suggestions. We also looked at the content of *problem descriptions* written by participants as a secondary measurement of task performance.

Behavioral Metrics. These were calculated from the tool's log data, including the total *time* spent on an analysis, the number of *pauses*, the duration of *pause time*, the number of forward *jumps*, and the number of *rewinds*. We considered a pause when a participant paused the video for 3 seconds or more, to avoid counting any unintentional actions. A forward jump was defined as a participant fast-forwarded the video by clicking on the timeline with more than 2 seconds apart from the current video time. Similarly, a rewind was counted when a participant went back in the video time by 2 seconds or more.

Subjective Perception. We relied on questionnaire responses to measure participants' *satisfaction*, *understanding*, and *trust* of AI Assistant. Specifically, satisfaction was measured by a three-item scale based on the After-Scenario Questionnaire [53] (Cronbach's alpha = 0.72). Understanding was

measured by two self-reported items: "I felt that I had a good understanding of how AI Assistant works" and "I felt that I had a good understanding of why AI Assistant detects UX problems" (Cronbach's alpha = 0.86). Trust was measured by a three-item scale (e.g. "I feel like I can count on AI Assistant to provide reliable suggestions for analysis of think-aloud sessions") adapted from "trust intention" in McKnight's framework on Trust [60, 61] (Cronbach's alpha = 0.72). All items were rated on a 7-point Likert scale. The Cronbach's alpha values, as indicated above, showed high internal consistency in the questionnaire responses.

4.5 Analysis Methods

For quantitative data, we computed descriptive statistics of the corresponding measures in Section 4.4 and performed mixed-effects regression models, which will be described in detail in Section 5.1. For qualitative data, the interview recordings were first transcribed into texts, and two researchers of the team performed thematic analysis on the texts independently and discussed the common themes that emerged from the texts. Finally, the themes were further discussed with two additional researchers of the team and consolidated into the key findings, which will be described in detail in Section 5.2.

5 RESULTS

We first present quantitative results from the experiment and then qualitative results from the interviews. We also annotate each subsection regarding the RQs that they primarily answer.

5.1 Quantitative Results

Statistical method overview. We conducted quantitative analysis on various dependant variables related to participants' task performance, behavioral patterns interacting with AI Assistant and subjective perception of the tool. We normalized the measurements that were likely impacted by either the video length or the number of ground-truth UX problems, detailed in each analysis below. For each dependant variable, we performed a separate mixed-effects regression with *explanations* (*w/* or *w/o*) and *synchronization* (*sync* or *async*) as the fixed effects and *participants* as the random effects. The regression model also included the interaction effect between explanations and synchronization. The regression model further included the *usability test videos* (website or coffee machine), participants' self-reported *UX experience*, *knowledge of AI* and general *trust in AI*, *age group*, and *gender* as control variables. All tests were performed with the *nlme* package in R. For each test, we checked for outliers with the dependant variable as outside 1.5 times the interquartile range (we will only mention below if a test had outliers identified and removed), and made sure there was no multicollinearity (all $VIF < 3$). Below we report the descriptive statistics (mean values and standard division) for each of the quantitative measures and significant results from the regression analysis.

Technical breakdowns happened for three participants: both tasks for P2 (*w/o explanations*) and P12 (*w/ explanations*), and one task (*async*) for P9 (*w/ explanations*), so we removed these five data points. On average, for the website video, participants spent 643 seconds (SD=48.5) and found 4.86 (SD=0.45) UX problems (ground-truth UX problems is 8; video length is 3.57 minutes). For the coffee machine video, participants spent 1331 seconds (SD=76.9) and found 11.9 (SD=1.04) UX problems (ground-truth UX problems is 20; video length is 11.37 minutes).

5.1.1 Task Performance (RQ1, RQ2). **Overall performance measurements.** We started by analyzing the measurements reflecting how well participants performed the tasks as described in Section 4.4, specifically the total numbers of UX problems identified, the Precision and Recall of their identified problems based on the ground truth problems. Given the difference between the two

Table 2. The mean and standard deviation (in parenthesis) of task performance measures: the number of UX problems identified by participants (N Problems) normalized by the number of ground-truth problems in each video, the precision and recall of the identified problems, and the length of problem description for each entry of UX problem (Desc. Len.)

		N Problems (normalized)	Precision	Recall	Desc. Len.
w/	*async*	0.69	.860	.560	11.4
		(0.09)	(.049)	(.055)	(2.2)
	sync	0.56	0.865	.475	10.2
		(0.06)	(.039)	(.047)	(1.9)
w/o	*async*	0.50	.939	.448	11.3
		(0.07)	(.037)	(.047)	(1.8)
	sync	0.66	.894	.589	13.5
		(0.08)	(.051)	(.068)	(2.1)

usability tasks (20 UX problems in the coffee machine video, 8 in the website video), we normalized the total number of UX problems found by the number of ground-truth problems in each video. Descriptive statistics, including means and standard deviations, are presented in Table 2 (Columns N Problems, Precision, and Recall).

We performed a mixed-effects regression model, as described above, on the number of identified problems (normalized by number of ground-truth problems), Precision, and Recall respectively. For the number of identified problems (normalized), we found a significant two-way interaction between the presence of explanations and synchronization ($\beta = -0.29, SE = 0.13, F(2, 18) = 4.54, p < 0.05$). Post-hoc analysis using *emmeans* package of R found the contrast between *async* and *sync* marginally significant for AI Assistant *w/o explanations* ($p = 0.10$)[1], but not significant for AI Assistant *w/ explanations* ($p = 0.22$), suggesting that the interactive effect was mainly caused by the difference of synchronization made for participants interacting with AI Assistant *w/o explanations*.

While we did not find any significant effect on Precision, we found the same significant two-way interaction between the presence of explanations and synchronization on Recall ($\beta = -0.23, SE = 0.09, F(2, 18) = 6.52, p = 0.02$; post-hoc analysis found the contrast between *async* and *sync* significant for AI *w/o explanations* ($p = 0.04$), but not significant for AI *w/ explanations* ($p = 0.17$)). No main effect was found. These two-way interactions on total number of problems and Recall are illustrated in Figure 3: when interacting with AI Assistant ***w/o explanations***, participants found significantly **more UX problems** from the ground truth, in the ***sync*** than the ***async*** condition. For those interacting with AI Assistant ***w/ explanations***, they did not show such difference in the *sync* than the *async* condition.

Overlaps with AI. We further examined how the AI influenced participants' analysis by the percentage of their identified UX problems overlapped with AI's suggestions, as divided by the total number of AI suggested problems (Pct. AI suggestions Confirmed), as well as overlap percentages over AI's false positive suggestions (Pct. AI FP Confirmed) and false negative suggestions (Pct. AI FN Found). Specifically, Pct. AI suggestions Confirmed indicates participants' propensity to confirm AI's suggestions. Pct. AI FP Confirmed indicates participants' tendency to be misled by AI's suggestions when there were no UX problems; and Pct. AI FN Found indicates participants'

[1]Given the relatively small sample size, we consider $p < 0.05$ as significant, and $0.05 \leq p < 0.10$ as marginally significant, following statistical convention [20]

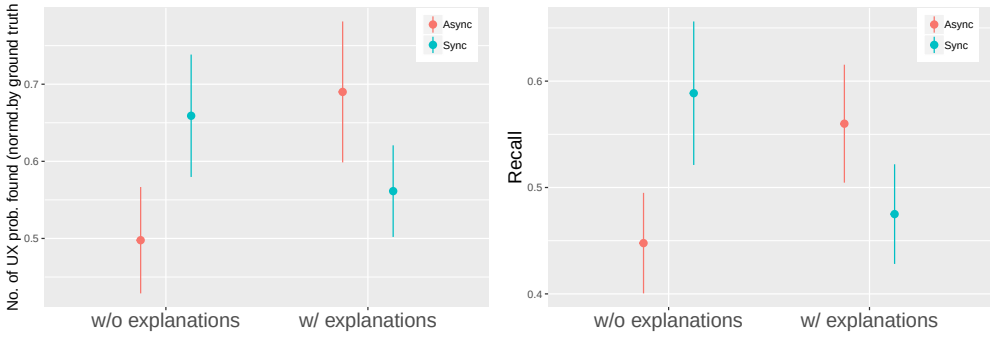

Fig. 3. Two-way interaction between explanations and synchronization on the number of UX problems found (normalized by the number of problems in ground truth), and Recall, as shown by means and standard divisions across conditions. All error bars represent +/- one standard error.

Table 3. The mean and standard deviation (in parenthesis) of measurements on overlaps with AI's suggestions and mistakes: the percentage of identified UX problems overlapped with AI's suggestions over the total number of AI suggested problems (Pct. AI suggestions Confirmed), overlap percentages over AI's true positive suggestions (Pct. AI TP Confirmed), overlap percentages over AI's false positive suggestions (Pct. AI FP Confirmed), and overlap percentages over AI's false negative suggestions (Pct. AI FN Found).

		Pct. AI suggestions Confirmed	**Pct. AI TP Confirmed**	**Pct. AI FP Confirmed**	**Pct. AI FN Found**
w/	*async*	53.2%	55.4%	21.7%	23.3%
		(6.3%)	(20.8%)	(9.3%)	(10.0%)
	sync	47.5%	60.5%	19.7%	12.1%
		(4.7%)	(20.1%)	(7.0%)	(5.1%)
w/o	*async*	40.6%	52.6%	3.0%	12.1%
		(4.2%)	(19.7%)	(3.0%)	(9.3%)
	sync	56.5%	62.0%	19.7%	21.2%
		(6.0%)	(24.9%)	(7.4%)	(10.3%)

ability to identify a correct UX problem when AI failed to make a suggestion. Descriptive statistics of these measurements are presented in Table 3.

We performed separate mixed-effects regression analysis as described earlier on each of the three measurements. We found the two-way interaction between the presence of explanations and synchronization to be significant for Pct. AI suggestions Confirmed ($\beta = -0.22, SE = 0.09, F(2, 18) = 6.94, p = 0.02$; post-hoc analysis found the contrast between *async* and *sync* significant for AI *w/o explanations* ($p = 0.02$), but not significant for AI *w/ explanations* ($p = 0.26$)) and marginally significant for Pct. AI FP Confirmed ($\beta = -0.18, SE = 0.10, F(2, 18) = 3.43, p = 0.08$; post-hoc analysis found the contrast between *async* and *sync* significant for AI *w/o explanations* ($p = 0.03$), but not significant for AI *w/ explanations* ($p = 0.76$)) . No main effect was found. As illustrated in Figure 4, these interaction effects show that for participants interacted with AI Assistant ***w/o explanations***, they accepted **more AI's suggestions** and **more false positive AI suggestions** in the ***sync*** than ***async*** condition. This difference, however, was not found for participants interacting with AI Assistant ***w/ explanations***. Figure 4 also suggests participants interacting with AI *w/*

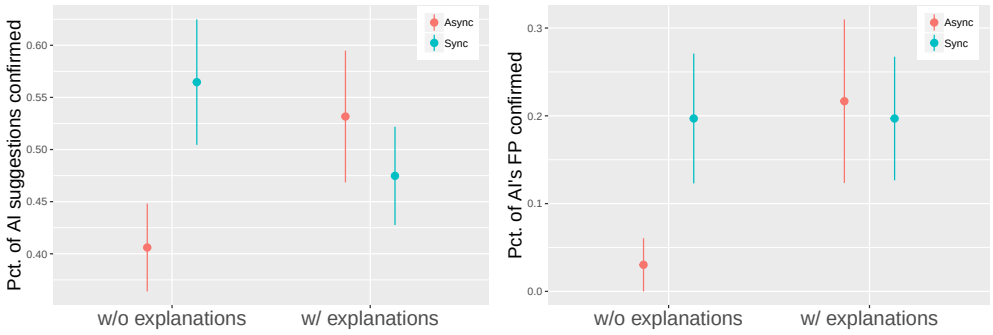

Fig. 4. Two-way interaction between explanations and synchronization on the percentage of AI suggestions confirmed, and percentage of AI's false negative suggestions confirmed, as shown by means and standard divisions across conditions. All error bars represent +/- one standard error.

explanations generally accepted more AI suggestions, in both *sync* and *async*, than the baseline condition (*w/o, async*).

What's more, Table 3 also suggests that AI's FPs and FNs had different effects on participants' analysis. In all conditions, the percentages of FPs leading to confirmed UX problems were low; this means that in most cases participants were able to recognize and reject AI's false suggestions. The percentages of FNs leading to confirmed UX problems were also low; this suggests that participants tended to neglect a UX problem if they were not reminded by the AI. These patterns reveal that in the context of identifying UX problems, FNs seem to be more harmful than FPs. The implication is that when designing a human-AI collaboration tool for UX evaluators, **it would be desirable to minimize FNs (i.e., the AI does not highlight the problem but it was actually there) over FPs (i.e., the AI highlights a problem but it was actually not there)** if trade-off between the two have to be made, for example by setting the classification threshold to have high recall.

Interestingly, after removing 3 outlier data points, we found that for Pct. AI FN Found, there is a marginally significant positive main effect of explanations ($\beta = 0.16, SE = 0.08, F(1, 15) = 4.18, p = 0.06$; M(*w explanations*)=13.3%; M(*w/o explanations*)=8.3%). This implies that participants interacted with AI Assistant ***w/ explanations*** were able to **find UX problems more effectively** even if AI failed to remind them.

5.1.2 Behavioral Patterns (RQ1, RQ2). We examined behavioral metrics that reflected how participants interacted with the tool, including the time spent analyzing a video, number of times they jumped forward in the video, rewound, paused, and the average time of each pause. Since the length of video might impact these behaviors, we normalized all metrics, except the average time per pause, by the length of the video (3.57 minutes for the website video and 11.37 minutes for the coffee machine video). That is, a normalized metric represents a participant' frequency of engagement in a corresponding behavior per every minute of a video. Table 4 shows descriptive statistics of these metrics.

We performed a separate mixed-effects regression on each of these behavioral metrics. We found a significant interaction effect between the presence of explanations and synchronization on the normalized total time spent ($\beta = -0.99, SE = 0.40, F(2, 18) = 6.03, p = 0.02$; post-hoc analysis found the contrast between *async* and *sync* significant for AI *w/o explanations* ($p = 0.01$), but not significant for AI *w/ explanations* ($p = 0.48$))), and a marginally significant effect on the average pause time (after removing an outlier) ($\beta = -7.12, SE = 4.05, F(2, 17) = 3.09, p < 0.10$; post-hoc analysis found

Table 4. The behavioral metrics reflecting participants' interactions with AI Assistant (with means and standard deviation in parenthesis), including the total time spent, number of jump forwards, rewinds, and pauses, all of which are normalized by the length of the video (in minutes); and average duration of a pause (in seconds).

		Total time (normalized)	N Jump Forward (normalized)	N Rewind (normalized)	N Pause (normalized)	M Pause time
w/	*async*	2.68	0.57	1.67	2.13	21.7
		(0.24)	(0.26)	(0.44)	(0.38)	(2.85)
	sync	2.21	0.16	1.25	1.61	18.1
		(0.23)	(0.09)	(0.43)	(0.30)	(2.34)
w/o	*async*	2.15	0.18	1.02	1.78	13.7
		(0.32)	(0.09)	(0.34)	(0.40)	(2.49)
	sync	2.83	0.30	1.51	1.70	17.6
		(0.41)	(0.09)	(0.43)	(0.32)	(2.71)

the contrast between *async* and *sync* marginally significant for AI *w/o explanations* ($p = 0.10$), but not significant for AI *w/ explanations* ($p = 0.38$))). As shown in Figure 5, these interaction effects are consistent with the trend of the identified UX problems, Recall, and percentage of AI suggestions confirmed described in the last subsections: When interacting with AI Assistant ***w/o explanations***, participants **spent more time and paused longer** in the *sync* than the *async* condition. Such a difference did not appear for participants interacting with AI assistant *w/ explanations*. Although not statistically significant, the same trend of interaction effect was observed for the number of forward jumps and rewinds as shown in Table 4.

Taken altogether, these behavioral patterns suggest that when interacting with AI Assistant ***w/o explanations***, participants appeared to be **more engaged** and examined the usability test videos **more actively** in the *sync* than the *async* condition. As a result, they found more UX problems in the *sync* than the *async* condition. For those interacting with AI Assistant ***w/ explanations***, their engagement and performance did not show significant difference between the *sync* and *async* conditions. Figure 5 suggests participants interacting with AI *w/ explanations* were generally more engaged (spent more time and paused longer) than the baseline condition *(w/o, async)*.

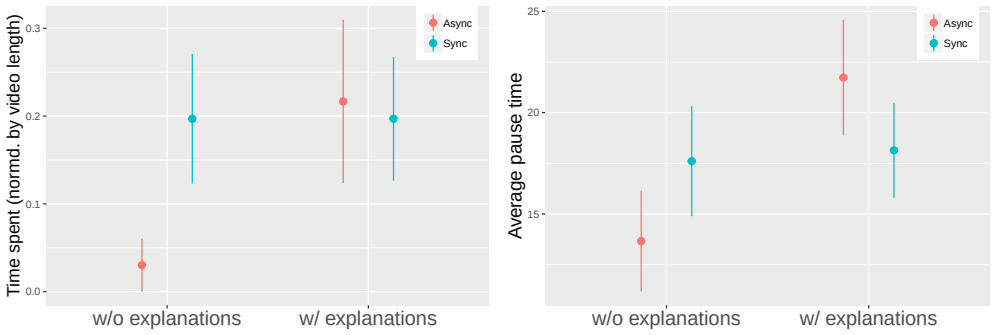

Fig. 5. Two-way interaction between explanations and synchronization on the time spent on analysis (normalized by video length, and average pause time, as shown by means and standard divisions across conditions. All error bars represent +/- one standard error.

Table 5. The subjective perceptions (e.g., satisfaction, understanding, and trust) of AI Assistant reported in the post-test questionnaire, with mean values and standard errors in parenthesis

		Satisfaction	Understanding	Trust
w/	*async*	5.53	5.90	4.80
		(0.23)	(0.22)	(0.40)
	sync	5.48	5.77	4.94
		(0.36)	(0.26)	(0.38)
w/o	*async*	4.88	4.64	4.55
		(0.42)	(0.47)	(0.40)
	sync	5.39	4.82	4.88
		(0.25)	(0.44)	(0.29)

5.1.3 Subjective Perceptions (RQ1-3). We analyzed the measurements of the subjective perceptions of the tool from questionnaire responses. After analyzing videos with each version of AI Assistant (*sync* or *async*), participants completed a questionnaire rating their *Satisfaction*, *Understanding* and *Trust* in AI Assistant, as described in Section 4.4. Table 5 shows the descriptive statistics of these questionnaire responses.

Separate mixed-effects regression models on the three subjective perceptions found a main effect of explanations on perceived Understanding ($\beta = 0.92, SE = 0.44, F(1, 15) = 4.35, p = 0.05$), suggesting that participants found the explanations helpful for them to better understand how AI Assistant worked. While not statistically significant, the measure of satisfaction exhibits a consistent trend with the results of task performance and behavioral patterns as indicated in previous subsections ($\beta = -0.55, SE = 0.47, F(2, 18) = 1.36, p = 0.26$). The main effect of explanations is also trending significant ($\beta = 0.68, SE = 0.52, F(2, 15) = 1.73, p = 0.20$). That is, for participants interacting with AI Assistant ***w/o explanations***, they felt **somewhat more satisfied** in the *sync* than the *async* condition. For participants interacting with AI Assistant ***w/ explanations***, they felt they **understood the AI better**, and **were somewhat more satisfied** with the tool, in both the *sync* and *async* conditions, than the baseline condition *(w/o, async)*.

It is worth noting that as controlled variables in regression models, we found a main effect of general Trust in AI in increasing Satisfaction ($\beta = 0.65, SE = 0.36, F(1, 15) = 0.09, p = 0.09$), Understanding ($\beta = 0.99, SE = 0.36, F(1, 15) = 7.93, p = 0.01$) and Trust ($\beta = 0.97, SE = 0.39, F(1, 15) = 6.09, p = 0.03$). This indicates that participants with a generally positive attitude towards AI also had more positive perceptions of AI Assistant.

In summary, we found that synchronization had a significant effect for participants interacting with AI Assistant *w/o explanations*: *sync* AI led to higher engagement (more time spent and longer pause), more acceptance of AI suggestions, better performance of finding UX problems, and somewhat higher satisfaction with the AI Assistant. Such difference from synchronization was not found for the group interacting AI Assistant *w/ explanations*, who were also found to have a better perceived understanding of AI Assistant, higher ability to detect usability problems from the video even if the AI Assistant failed to remind them. . The latter group also exhibited a tendency of higher engagement and more acceptance of AI suggestions compared to the baseline condition *(w/o, async)*.

5.2 Qualitative Results

Based on the quantitative results above, we first discuss insights from qualitative data to further unpack the benefits of AI explanations and synchronization, and how they might have worked together to create the interactive effect on participants' engagement and performance. We will

also discuss themes on the general perception of AI Assistant to support UX evaluators to perform usability test video analysis (Section 5.2.3) and suggest ways to improve the design (Section 5.2.4). We will highlight the themes identified from the thematic analysis in **bold** and support them with selected quotes. To help readers contextualize the quotes, we annotated them with their corresponding Explanation conditions *(w vs. w/o)* since Explanation was a between-subjects factor and each participant experienced only one of the two conditions. In contrast, since Synchronization was a within-subject factor and all participants experienced two conditions *(sync vs. async)*, we only annotated ones that were clear which synchronization condition they referred to.

5.2.1 Benefits of Explanations (RQ1). First, we found further support consistent with the quantitative results that having access to AI's explanations led to a perceived **better understanding** of how the AI works.

> "*I felt that I could rely on it more than half of the time. It detected what the participant said and did. By combing these two factors, I felt it often generated reasonable suggestions.-P18, w/*"

Despite a lack of significant survey results on user trust, we found some evidence that, by having explanations, the **improved user understanding can potentially improve trust** in AI Assistant. This is illustrated by the lower frequency of mentioning a lack of trust among participants who used the AI *with* explanations (3 out of 12) than those who used the AI *without* explanations (9 out of 12), for example:

> "*Maybe after I work with the AI for a while and really feel it helps me identify problems, I will build trust with it. But for now, since I feel I don't know enough about the AI and I don't know how it works, I am not sure I can 100% trust it.-P23, w/o*"

Second, participants felt the explanations had **utility for UX evaluation tasks**, which could help them make more informed judgments of UX problems:

> "*I trust myself more than the AI. For example, while the AI thought the user encountered an issue when searching for the instruction picture, I think it was just a typical searching process and was not necessarily an issue. But I think it is valuable to review what the AI says and why and then make my own decision.-P23, w/o*"

Consistent with the observations that participants exhibited better ability to detect UX problems and higher engagement when explanations were presented, many commented on **attending to the explanation contents**:

> "*I like the tool was able to catch some very detailed stuff because I was not able to divide my attention to multiple parts of the whole video like both the participant face video and the screen capture.-P17, w/*"
>
> "*I was bad at detecting the uncertainty in the user's tone and hoped that the tool was able to pick that up.-P20, w/o*"

This utilization of explanations content is also demonstrated by participants' tendency to incorporate the content in AI's explanations in their submitted UX problem descriptions, as discussed in the quantitative analysis shown in the last column of Table 3.

The above qualitative findings highlight the benefits of providing AI explanations in assisting usability video analysis. Not only did they help users better understand the AI, but also provided additional task support to help identify usability problems. These benefits may explain participants' improved analysis, especially their ability to identify more false-negative problems, even when the AI did not explicitly suggest. Interestingly, we did not find a compromise of human agency by providing explanations, as suggested by the notion of human-AI agency divide [49, 50], but instead found improved human engagement in analyzing the video as compared to the baseline condition.

5.2.2 Benefits of Synchronization (RQ2). A key finding of our quantitative analyses was that when explanations were not present, participants better engaged with the videos and performed significantly better when the AI suggestions were provided *synchronously* than asynchronously. Our qualitative data revealed four main reasons how the *sync* AI helped participants.

First, participants felt that the *sync* AI allowed them to better perform **independent analysis**, and not be biased by the amount and places of UX problems in a video, as illustrated by the following quote from a participant interacting with AI Assistant .

> "*I don't prefer it (the async AI) showing me many indications of problems before I make my own judgment. I don't want to be biased.-P24, (w/o, async)*"

Second, compared to the *async* AI, which showed the AI's suggestions all at once since the beginning of their analysis, the *sync* AI was **less overwhelming**.

> "*Async [AI] is a bit overwhelming with the problems revealed all at once. It could bias me before I am able to look into the video and have my judgments.-P20, (w/, async)*"

Relatedly, participants found the *sync* AI allowed them to **better focus** on their current analysis without being distracted by AI's suggestions for the forthcoming portion of the video.

> "*[With] the sync [AI] I can pay more attention to what's going on right now. But the async [AI] also tells you the next usability problem that is going to happen. In that case, I found myself tend to focus more on what's going to happen next instead of finding the issue that is happening now. So it distracted me from my current analysis.-P19, (w/o, sync)*"

Last but not least, the *sync* AI created a stronger sense of **social presence**, with the perception of analyzing the video with another "evaluator" simultaneously.

> "*It feels like another person is doing this [analysis] with me.-P1, (w/, sync)*"

Although the above results suggest *the async AI allowed lower human agency*, participants also pointed out two potential benefits. First, the *async* AI allowed them to get an **overview** of the distribution of the potential UX problems in the video. Second, the *async* AI resulted in **less frequent visual updates** on the interface than the *sync* AI, therefore was less demanding on their attention and cognitive resources.

> "*I really like the AI highlighted all the problems at the beginning so that way I could pay attention to certain parts of the video. In that way, I could get things done much faster.-P15, (w/o, async)*"

The observations in the last two sub-sections can help interpret the interactive effect between synchronization and explanations. For the AI Agent *w/o explanations*, participants repeatedly expressed a level of mistrust in its capability. The synchronous AI, which progressively reveals its suggestions as the evaluator examines the video, validating their analysis rather than dictating the job, might have been a better match than the asynchronous AI to support a desired level of human agency. As a result, UX evaluators had better engagement in analyzing the video and better analysis outcomes.

This mismatch between the asynchronous AI and UX evaluators' desired human agency might have been mitigated with the explanations provided by the AI Assistant, which could have been perceived to be more capable and trustworthy. Perhaps more importantly, the content of explanations provided additional information and utility for evaluators to analyze the video, which could have helped them better identify usability problems and stay more engaged regardless of the synchronization, as the quantitative results showed.

5.2.3 Perceived Values of AI Assistant for Usability Video Analysis (RQ3). To answer RQ3, we summarize the key values that participants saw from the WoZ AI Assistant to inform effective forms of human-AI collaboration for usability video analysis. First, participants appreciated AI to

provide a **second opinion to verify** their analysis, which UX practitioners often desire but would be expensive to obtain from another human evaluator since most times they perform usability analysis alone [25, 27].

> "*I can compare what I think is a problem to what the AI thinks is a problem. It gave me a second opinion without calling for a second designer for help.-P11, w/o*"
>
> "*Oftentimes one researcher has to cover several projects, so it's certainly nice to have a second opinion and also it can correct me and identify something I miss.-P23, w/o*"

Interestingly, some felt that the AI Assistant was like a "junior colleague" or *"a new intern who needs to be monitored and costs time and energy"* (P17). Given the performance limitation of AI and subjectivity of usability video analysis, this could be a proper mental model to design for effective human-AI collaboration in this context. We may argue that *sync* AI, without dominating the analysis, is a better fit for such a mental model. It would also be interesting to explore retrospective AI that reveals its suggestions after the evaluators' judgment.

Second, participants desired for the AI to **accelerate their analysis** by suggesting where to focus on in the video, especially when the video is long or when UX evaluators work under time constraints.

> "*The AI might speed up the whole evaluation, especially with time constraints. I can use it, to some degree, to accelerate my analysis by finding which parts I should go a little deeper.-P20, w/o*"

However, our quantitative analysis revealed the risk of being misled by AI's false negatives by following AI's suggestions to navigate the video. Future work should balance the efficiency aspect and users' engagement with analyzing video content.

Third, the AI could **complement** participants' analysis by identifying issues that they might otherwise miss. As discussed in Sec. 5.2.1, many participants found the explanations particularly helpful for directing their attention to details that they did not recognize.

> "*It gave me suggestions before I found issues and also reminds me of the places that I did not recognize the issue.-P12, w/*"

Finally, some participants used AI's suggestions, such as the timeline feature, as **anchor for further analysis** or to know where to replay the video.

> "*When I went back to watch the video for the second time, I knew where to pay attention to with the visualization of the problems.-P15, w/o*"
>
> "*When I double-checked my analysis by playing the video back and forth, it helped me locate the portions where usability problems might exist.-P13, w/*"

5.2.4 Ways to Improve the Collaboration between UX Evaluators and AI (RQ1-3). Participants also expressed ways in which the AI could be improved to better assist them with their analysis.

Explanations. The current AI Assistant describes five behavioral features in the visual and audio of a usability test video to explain the AI's suggestions. However, participants had different individual preferences for the importance of these features, and desired for ways to **personalize the rank and types** of features provided.

> "*I would rank the five features in this way: what the user said, what the user did, sentiment, tone, and speech rate because i felt tone and speech rate were not as important as the other three.-P13, w/*"
>
> " *I would suggest making speech rate optional. It's not critical for decision making.-P24, w/o*"

Participants also had **different preferences for when** the explanations should be shown. While some suggested the explanations to be shown only when they request them, others prefer seeing the explanations all the time instead of just the times when the AI detects problems.

> "*I hope the AI can show the information when I need it instead of providing information to me all the time.-P13, w/*"
>
> "*Instead of presenting this dialogue of information only during the time when the AI thinks there is a usability problem, maybe just show it through the video, which would be easier for me to analyze the problems.-P16, w/*"

Synchronization. Participants suggested three alternative designs to consider the timing to reveal AI suggestions. The first approach is to allow UX evaluators to analyze the video by themselves first, and then the AI reveals its suggestions to help verify judgments in the **second pass**.

> "*I prefer analyzing the video by myself first and then take a second look at it, and that is when the AI can show me the result because I do want to get a different perspective on what's going on in the video.-P24, w/o*"

Second, some participants suggested retrospective AI, by revealing its suggestions **immediately after the UX evaluator indicates their judgment**, for example, by showing agreement or flagging missed problems. Such designs would further avoid biasing the UX evaluator's judgments.

> "*I think it would be more helpful if the AI can analyze with me at the exact same time. For example, it shows me its inference at the exact same time when I find an issue.-P12, w/*"

The third approach is to let the AI analyze a video asynchronously but **flag suggestions that the AI is uncertain about**. In this way, UX evaluators could save time on the problems that AI is confident about and focus their attentions on the ones that AI would likely make mistakes.

Additional Features and Functions. Participants also suggested additional features and functions. First, participants expressed the desire to know **richer information about AI**, such as the accuracy of the AI and the principles that the AI uses. Moreover, they wanted to see more information about UX problems, such as the type, frequency, and severity of the suggested problems, and perhaps even recommendations for fixing the problems.

> "*I don't know the authenticity of the AI, such as its accuracy...It's nice to have many types of principles, not only Nielsen's, for the AI to apply and have an option to select a set of principles for the AI, which might generate different results.-P11, w/o*"
>
> "*Instead of presenting this dialogue of information only during the time when the AI thinks there is a usability problem, maybe just show it through the video, which would be easier for me to analyze the problems.-P16, w/*"

Second, they hoped that the AI could adapt its inferences by **learning from human guidance**. For example, UX evaluators could analyze a few usability test sessions first, and then the AI learns their analysis and automatically analyzes the remaining sessions. A different approach is to let UX evaluators review and edit the AI's suggestions for the AI to learn from. These active learning approaches could allow the AI to be better customized to detect UX problems in specific technology domains. Furthermore, they suggested another way to improve AI is to allow for choosing different UX design principles, other than Nielsen's heuristics [63] as used in this work, for the AI to consider.

Last, participants also hoped the AI could generate **a report of UX problems**. For example, the AI could generate the description for each problem, which UX evaluators could either directly use or edit further if needed when they create their UX analysis report.

> "*The AI may provide some problem descriptions, so we just need to check the box to select or edit instead of having to type the descriptions all by ourselves. It might make the analysis much faster.-P12, w/*"

6 DISCUSSION

To summarize, we started from the baseline design of AI Assistant, which was the asynchronous AI without explanations *(w/o, async)*, and extended it into three additional designs of AI by changing two factors–explanations and synchronization. We found that:

- **Explanations**: When the AI Assistant provided explanations, participants performed equally well regardless of synchronization, and exhibited higher engagement and more acceptance of AI suggestions compared to the baseline AI *(w/o, async)*. Explanations also improved participants' self-reported understanding of the AI, ability to detect usability problems even if the AI Assistant failed to remind them .
- **Synchronization**: When without explanations, *Synchronous* AI improved UX evaluators' performance and engagement with AI's suggestions more than the baseline AI *(w/o, async)*.

Next, we first elaborate on these findings and their implications for designing AI explanations and synchronization, and then discuss high-level takeaways regarding human-AI collaboration and AI for UX work.

6.1 Design Implications: Explanations and Synchronization

While our quantitative results suggest both explanations and synchronization could benefit human-AI collaboration in the context of UX evaluation, our qualitative data further reveal the underlying reasons and suggest design implications

Explanations. Consistent with prior work [12, 17, 47], we found that providing explanations could robustly improve users' perceived understanding of AI. More importantly, participants appreciated the additional support enabled by explanations, to better direct their attention to the parts of the video likely associated with UX problems, and to remind them of the indicators of UX problems that they might have otherwise missed. Some participants even wished to have constant access to the explanation of the AI's input features, suggesting that they did not merely view these features as supporting AI's suggestions, but also utilized them in their own analysis of the videos.

Our WoZ design of AI explanations provided two pieces of information to support UX evaluation: violations of design principles or heuristics [63] (rule-based explanations), and behavioral features of the test subjects that are potential indicators of UX problems [24] (feature-based explanations). Future work could explore technical approaches to generating such explanations in a robust way.

Further, our qualitative data reveal varying individual preferences for the types and ranks of AI's features to be shown, and a desire to customize the explanations. It is worth exploring ways to balance two perspectives of using explanations: as *justification* of AI's decisions to better understand the AI and as *additional information* to better support the UX evaluation task. Current XAI methods usually adopt the former perspective. A fixed set rules or features are shown as explanations based on how the AI actually arrives at its decisions. There is a growing criticism that this algorithm-centric view may miss great opportunities to support a primary motivation why people seek explanations [62]: a joint sense-making process where both the explainer and explainee supply information or constraints to build common ground. We believe knowledge-rich domains where experts and AI collaborate, such as the usability video analysis task we studied, provide interesting test grounds to explore novel interactive explanation techniques that allow experts to set preferences or constraints, even provide feedback or critique, for AI to improve its explanations so as to better support the experts' tasks.

Lastly, we should point out that recent works warned that explanations could lead to unwarranted trust [9, 29, 72, 95] and over-reliance on AI, if a user simply associates explanability with AI capability without engaging analytically with the model behaviors. While our results did not suggest this is a salient problem, future work introducing real, imperfect AI technologies should be mindful

about such a risk. For example, recent studies suggested adaptive explanations—only providing explanations for suggestions that the AI is confident about [8, 95].

Synchronization. In current AI-mediated decision support systems, it is common to adopt an asynchronous user mental model, in which the AI finishes the analysis first and presents its analysis to domain experts (e.g., [26]). Our study, however, reveals valuable benefits of providing synchronous AI support, unfolding in real time as the user performs the task. With the synchronous AI, our participants particularly welcomed their ability to retain high human agency, not get biased or distracted by AI's suggestions beyond the portion of the video that they have analyzed, and get a second opinion to verify or complement their judgment, which is often lacking in UX practice [25, 27] but important for overcoming the "evaluator effect" [38].

Furthermore, the synchronous AI also helped to create a stronger "social presence," the feeling of working together with a colleague, than the asynchronous AI. In CSCW literature, studies repeatedly found that workers exhibit higher engagement, creativity and social interactions in synchronous than asynchronous collaboration contexts [11, 78], while asynchronous collaboration may incite social loafing, with which an individual exerts less effort working with others. This risk could possibly extend to human-AI collaboration, as we observed less engagement and lower task performance in the baseline asynchronous condition than the synchronous conditions.

It is interesting that the comparative disadvantages of the asynchronous AI Assistant were mitigated if it provided explanations. We suspect that it may relate to users' mental model of the capability of AI Assistant. In the qualitative data, participants repeatedly mentioned that they did not place high confidence on the AI's capability to detect all UX problems in the videos, seeing it as a "junior colleague," and preferred to rely on themselves and only utilized the AI's suggestions as verification or reminders. Thus, without dictating the analysis or distracting from their own judgment, the synchronous AI was a better fit for such a mental model than the asynchronous AI, which pushed all its judgements to them before they even started their own analysis. When the explanations panel was added (*AI w/ explanations*), participants viewed the explanations as an additional utility feature to guide them better navigate and observe the actions in the videos, which may explain their equal engagement in both the synch and asynch AI conditions.

It is worth noting that the takeaway message is not that synchronous AI should always be preferred over asynchronous AI. Instead, this design decision should depend on the AI agency and the type of support that the AI can provide. It was the mismatch between the collaboration timing, where the AI dictates the analysis in the asynchronous condition, and the AI's limited range of support (i.e., without explanations) that led to the undesirable outcome. In knowledge-rich domains, such as UX evaluation, it is not uncommon that users place more faith in their domain expertise and judgments than the current AI and prefer to have a synchronous AI assistant. However, as users grow their trust in the AI, they might enjoy an asynchronous AI assistant that can provide an overview of the problems in the entire video as indicated in the last paragraph in Sec 5.2.2. Alternatively, users might prefer the flexibility to toggle between synchronous and asynchronous collaboration with AI.

6.2 Human Agency and Timing of Human-AI Collaboration

We would like to extend the point of designing for different Human-AI agency divide and matching the timing of Human-AI collaboration further. First, providing explanations is not the only way to increase the level of AI assistance, in terms of the quality and scope of support it can provide. Lai et al.'s [49, 50] work suggested another type of information to increase AI agency–performance information to give users confidence and relieve them from the cognitive effort in judging the AI's recommendations. We found similar suggestions from participants to see richer information about the AI's performance and confidence as indicated in Sec. 5.2.4. The implication is that designers need

to consider more nuanced designs for the timing factor of human-AI collaboration, for example, by allowing the AI to work asynchronously for high-confidence cases but synchronously for low-confidence cases.

There are several other frameworks on different levels of AI assistance for both broader automated systems [56, 66] and specific AI applications [51, 94]. Mackeprang et al. suggested a 10-level framework for different configurations of machine assistance, ranging from offering no assistance, to offering a set of suggestions, to automatically confirming with different ways for humans to intervene, to acting autonomously [56]. Our current designs of AI Assistant were at the lower end of this automation spectrum, with the AI only offering suggestions, so the synchronous collaboration was more suitable than the asynchronous one. In the future, if the AI becomes more competent and could be confidently tasked with auto-detection and report-generation for usability video analysis, the timing factor should be reconsidered accordingly. For example, some participants suggested letting the AI perform the analysis asynchronously and then flag the decisions that it is uncertain about for human evaluators to double check.

Last but not least, AI agency and human agency are not necessarily a dichotomous trade-off. On the one hand, we found some evidence that providing explanations improved participants' perceived AI capability, even trust. On the other hand, we did not find a compromise of human agency. To the contrary, explanations seemed to have "slowed down" the evaluators and improved their ability to detect usability problems even if the AI failed to suggest. Recent work by Shneiderman [79] highlights that human-AI agency is not necessarily a dichotomous trade-off. Instead reliable and safe AI design should strive for both high AI and high human agency to exercise effective control. Our results suggest that providing explanation could be a possible way to achieve this goal.

6.3 AI for UX Evaluation

Our work provides empirical insights into UX professionals' attitudes, needs and preferences for AI technologies to support UX evaluation work. On the one hand, they were positive about the idea of having AI support to accelerate and complement their work, especially given that UX professionals often analyze a test session alone under time pressure [25, 27, 58]. They especially welcomed the idea of relieving them from tedious or mechanic part of their tasks, such as maintaining close attention to lengthy usability test videos or writing up usability reports.

On the other hand, participants expressed doubt about AI's capabilities to perform usability analysis. One possible reason is that detecting usability problems requires the understanding of the test subject's various behavioral signals and the contexts of the test products and tasks, which was viewed by many participants as too complicated for the AI to fully comprehend.

As a result, participants suggested guiding or training the AI to be more similar to how they perform usability analysis. This suggestion hints that *adaptability* or *customizability* might be a promising area of technical and design innovation for AI-Assisted UX evaluation. Toward this goal, the fields of *active learning* [77] and *interactive machine learning* [2, 23] provide examples of techniques and systems that allow domain experts without ML expertise to train ML models by supplying examples and tuning the results with interactive interfaces. For example, *explainable active learning* [29, 83] elicits experts' feedback on AI's explanations and incorporates it to train an AI better aligned with how experts make judgments in the task domain. This seems to be a viable approach, as our participants expressed a strong desire to tune the AI, and UX practitioners in general tend to accumulate rich expertise in recognizing the violations of design principles or heuristics that are still too abstract for the AI to pick up from behavioral data alone.

Meanwhile, one should be cautious about the common pitfalls of active or interactive machine learning systems [2, 91], one of which is overfitting, i.e., an AI that fits precisely with expert-provided examples or rules but does not generalize. This is especially problematic as a primary

benefit of AI Assistant recognized by our participants was providing a second perspective on their analysis. However, an overfitted AI would reinforce their confirmation bias and further trap them in the "evaluator effect" [38]. One possible remedy is to explore ensemble approaches with diverse models trained by multiple UX professionals to jointly identify UX problems in usability test videos.

6.4 Limitations and Future Work

WoZ. While we discussed technical feasibility of implementing different pieces of information in the WoZ AI's explanations in Sec. 3.3.2, it remains an open question of how to implement such an AI that could generate explanations as good as the WoZ AI. The implication is two-fold. On the one hand, if UX evaluators tend to include AI's explanations into their personal justification, then poor quality or invalidate explanations from such a practical AI could reduce their task performance. On the other hand, we might expect to see lower satisfaction and trust in the AI among participants, who eventually would disengage with it.

Our study found that UX evaluators wanted to access more information about AI, such as accuracy and uncertainty. While exposing such information to users might help them understand AI's inferences, it might have downside too. For example, Lim and Dey found that participants could lose their trust in an intelligent agent that shows high uncertainty even if it has a high intelligibility (e.g., being able to explain its reasoning well) [55]. However, their study was conducted with participants using context-aware applications. It remains unknown how exposing uncertainty and accuracy of AI to UX evaluators would affect their collaboration with the AI.

We simulated the AI with reasonable precision and recall to match the expectation of an imperfect AI in practice. Recent research has suggested that the precision and recall of an AI might affect its users' perceptions [47]. Therefore, it remains an open question of how accessing precision and recall of the AI might affect UX evaluators' collaboration with the AI.

In-the-Wild Studies. As a controlled lab study, we could only include a limited set of UX evaluators and usability test videos. However, the diversity of the products on the market and the diverse backgrounds and experiences of UX evaluators might also play a role in how they use and perceive the AI when they analyze usability test videos. More in-the-wild studies with a broader set of products and UX evaluators are needed to further validate the findings.

Task Domains Beyond Usability Video Analysis. We have studied the effects of explanations and synchronization on human-AI collaboration in the context of UX analysis with tasks of analyzing usability test videos. One should be cautious about generalizing the findings to other task domains. The tasks used in our studies (i.e, identifying usability problems) require expertise in UX research. As a result, our participants were domain experts. Our research and prior work (e.g., [26, 80]) suggested that expert users tended to be confident in their judgements and might have adopted different strategies in human-AI collaboration compared to the crowdworkers or non-experters widely used in human-AI collaboration research. In addition to expertise requirements, the stake of tasks could be another factor affecting the generalization of the findings. While our tasks may be of higher stake than many crowdsourced tasks, much higher stake tasks (e.g., cancer diagnosis) exist. Thus, although our findings show the positive benefits of explanations and synchronization in fostering an engaging human-AI collaboration, more research is warrant to validate and extend our findings in other task domains (e.g., tasks requiring expertise vs. tasks without special expertise requirement; high-stake vs low-stake tasks).

7 CONCLUSION

We have studied how two factors of human-AI collaboration—explanations and synchronization—would affect the performance and perception of UX evaluators in the context of analyzing usability test videos. We iteratively designed a tool—AI Assistant—with four versions of UIs corresponding

to the two levels of explanations (with/without) and synchronization (sync/async). We conducted a mixed-methods study with 24 UX evaluators identifying UX problems from usability test videos with AI Assistant.

Our quantitative results show that both explanations and synchronization have positive effects on UX evaluators' performance (e.g., the number of identified UX problems) and engagement (e.g., time spent analyzing the video). Specifically, AI with explanations, regardless of being presented synchronously or asynchronously, helped to improve UX evaluators' performance of and engagement in their analysis, as well as perception of the tool (e.g., understanding and satisfaction), compared to the baseline AI (i.e., asynchronous AI without explanations); when without explanations, synchronous AI improved UX evaluators' performance and engagement more than the asynchronous baseline AI.

Our qualitative results further reveal the ways how explanations and synchronization helped UX evaluators. Specifically, explanations helped them better understand how the AI works, increased their trust in AI, and provided them with additional utility for UX evaluation tasks; and synchronization helped them better perform independent analysis, feel less overwhelmed by the AI's suggestions, and create a stronger sense of "social presence"—the feeling of working together with a "colleague."

Based on our findings, we have shown ways to improve the design of explanations and synchronization aspects of human-AI collaboration, such as matching the AI's agency with the type of support it can provide (i.e., AI's capability), balancing between human agency and AI agency, and supporting the adaptability and customizability by allowing UX practitioners to guide the AI, such as by offering their expertise. Finally, we have presented the design implications for AI agency and timing of human-AI collaboration in general and for AI-Assisted tools for UX evaluation.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[3] Morten Sieker Andreasen, Henrik Villemann Nielsen, Simon Ormholt Schrøder, and Jan Stage. 2007. What happened to remote usability testing? An empirical study of three methods. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1405–1414.

[4] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*. 5803–5812.

[5] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. 2020. OpenCrowd: A Human-AI Collaborative Approach for Finding Social Influencers via Open-Ended Answers Aggregation. In *Proceedings of The Web Conference 2020*. 1851–1862.

[6] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[7] Sriram Karthik Badam and Niklas Elmqvist. 2014. Polychrome: A cross-device framework for collaborative web visualization. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 109–118.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[10] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[11] Jeremy Birnholtz and Steven Ibara. 2012. Tracking changes in collaborative writing: edits, visibility and group maintenance. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 809–818.
[12] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.
[13] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[15] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019), 832.
[16] José C Castillo, H Rex Hartson, and Deborah Hix. 1998. Remote usability evaluation: can users report their own critical incidents?. In *CHI 98 conference summary on Human factors in computing systems*. 253–254.
[17] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 559.
[18] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.
[19] Kristin A Cook and James J Thomas. 2005. *Illuminating the path: The research and development agenda for visual analytics*. National Visualization and Analytics Ctr.
[20] Duncan Cramer and Dennis Laurence Howitt. 2004. *The Sage dictionary of statistics: a practical resource for students in the social sciences*. Sage.
[21] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.
[22] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data*. the MIT Press.
[23] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.
[24] Mingming Fan, Jinglan Lin, Christina Chung, and Khai N. Truong. 2019. Concurrent Think-Aloud Verbalizations and Usability Problems. *ACM Transactions on Computer-Human Interaction* 26, 5, Article 28 (July 2019), 35 pages. https://doi.org/10.1145/3325281
[25] Mingming Fan, Serina Shi, and Khai N Truong. 2020. Practices and Challenges of Using Think-Aloud Protocols in Industry. *Journal of Usability Studies* 15, 4 (2020).
[26] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N Truong. 2020. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 343–352.
[27] Asbjørn Følstad, Effie Law, and Kasper Hornbæk. 2012. Analysis in practical usability evaluation: a survey study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2127–2136.
[28] Jill Gerhardt-Powals. 1996. Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction* 8, 2 (1996), 189–211.
[29] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2021).
[30] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.
[31] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[32] Julián Grigera, Alejandra Garrido, José Matías Rivero, and Gustavo Rossi. 2017. Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies* 97 (2017), 129–148.
[33] Nielson Norman Group. [n.d.]. 10 Heuristics for User Interface Design: Article by Jakob Nielsen. https://www.nngroup.com/articles/ten-usability-heuristics/
[34] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[35] Patrick Harms. 2019. Automated usability evaluation of virtual reality applications. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–36.

[36] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1029–1038.

[37] Christian Herbst. [n.d.]. Christian's Python Library :: A Python library for voice analysis. https://homepage.univie.ac.at/christian.herbst/python/namespacepraat_util.html

[38] Morten Hertzum and Niels Ebbe Jacobsen. 2001. The evaluator effect: A chilling fact about usability evaluation methods. *International journal of human-computer interaction* 13, 4 (2001), 421–443.

[39] Michael Hind, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2019. TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 123–129.

[40] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[41] iMotions. 2021. iMotions. https://imotions.com/biosensor/fea-facial-expression-analysis/.

[42] Petra Isenberg, Niklas Elmqvist, Jean Scholtz, Daniel Cernea, Kwan-Liu Ma, and Hans Hagen. 2011. Collaborative visualization: definition, challenges, and research agenda. *Information Visualization* 10, 4 (2011), 310–326.

[43] Petra Isenberg, Danyel Fisher, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. 2010. An exploratory study of co-located collaborative visual analytics around a tabletop display. In *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 179–186.

[44] JongWook Jeong, NeungHoe Kim, and Hoh Peter In. 2020. Detecting usability problems in mobile applications on the basis of dissimilarity in user behavior. *International Journal of Human-Computer Studies* 139 (2020), 102364.

[45] Robert Johansen. 1988. *Groupware: Computer support for business teams*. The Free Press.

[46] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.

[47] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[48] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*.

[49] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[50] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[51] Doris Jung-Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *Data Engineering* (2019), 58.

[52] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[53] JR Lewis. 1995. Computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1 (1995), 57–78.

[54] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[55] Brian Y Lim and Anind K Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*. 415–424.

[56] Maximilian Mackeprang, Claudia Müller-Birn, and Maximilian Timo Stauss. 2019. Discovering the Sweet Spot of Human-Computer Configurations: A Case Study in Information Extraction. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[57] N. Mahyar and M. Tory. 2014. Supporting Communication and Coordination in Collaborative Sensemaking. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1633–1642. https://doi.org/10.1109/TVCG.2014.2346573

[58] Sharon McDonald, Helen M Edwards, and Tingting Zhao. 2012. Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication* 55, 1 (2012), 2–19.

[59] Will McGrath, Brian Bowman, David McCallum, Juan David Hincapié-Ramos, Niklas Elmqvist, and Pourang Irani. 2012. Branch-explore-merge: facilitating real-time revision control in collaborative visual exploration. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*. ACM, 235–244.

[60] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[61] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.

[62] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[63] Jakob Nielsen. 2005. Ten usability heuristics. https://www.nngroup.com/articles/ten-usability-heuristics

[64] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[65] Asil Oztekin, Dursun Delen, Ali Turkyilmaz, and Selim Zaim. 2013. A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems* 56 (2013), 63–73.

[66] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[67] Fabio Paternò, Antonio Giovanni Schiavone, and Antonio Conti. 2017. Customizable automatic detection of bad usability smells in mobile accessed web applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–11.

[68] Sharoda A. Paul and Madhu C. Reddy. 2010. Understanding Together: Sensemaking in Collaborative Information Seeking. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 321–330. https://doi.org/10.1145/1718918.1718976

[69] Lloyd R Peterson. 1969. Concurrent verbal activity. *Psychological Review* 76, 4 (1969), 376.

[70] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review* 106, 4 (Oct. 1999), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

[71] Pedro Ponce, David Balderas, Therese Peffer, and Arturo Molina. 2018. Deep learning for automatic usability evaluations based on images: A case study of the usability heuristics of thermostats. *Energy and Buildings* 163 (2018), 111–120.

[72] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[73] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. 93–100.

[74] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[75] Ali Sarvghad and Melanie Tory. 2015. Exploiting analysis history to support collaborative data analysis. In *Proceedings of Graphics Interface Conference*. Canadian Information Processing Society, 123–130.

[76] Ali Sarvghad, Melanie Tory, and Narges Mahyar. 2016. Visualizing dimension coverage to support exploratory analysis. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 21–30.

[77] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[78] Ashraf I Shirani, Mohammed HA Tafti, and John F Affisco. 1999. Task and technology fit: a comparison of two technologies for synchronous and asynchronous group communication. *Information & Management* 36, 3 (1999), 139–150.

[79] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[80] Ehsan Jahangirzadeh Soure, Emily Kuang, Mingming Fan, and Jian Zhao. 2021. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Transactions on Visualization and Computer Graphics* (2021), 1–1. https://doi.org/10.1109/TVCG.2021.3114822

[81] Siegfried Ludwig Sporer and Barbara Schwandt. 2006. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 20, 4 (2006), 421–446.

[82] Anselm Strauss. 1985. Work and the division of labor. *Sociological quarterly* 26, 1 (1985), 1–19.

[83] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 239–245.

[84] Matthew Tobiasz, Petra Isenberg, and Sheelagh Carpendale. 2009. Lark: Coordinating co-located collaboration with information visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1065–1072.

[85] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729* (2014).

[86] Fernanda B Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics* 13, 6 (2007).

[87] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[88] Martin Wattenberg and Jesse Kriss. 2006. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics* 12, 4 (2006), 549–557.

[89] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. *arXiv preprint arXiv:2005.00582* (2020).

[90] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: Structured support for collaborative visual analysis. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 3131–3140. https://doi.org/10.1145/1978942.1979407

[91] Tongshuang Wu, Daniel S Weld, and Jeffrey Heer. 2019. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 4 (2019), 1–27.

[92] Shenyu Xu, Chris Bryan, Jianping Kelvin Li, Jian Zhao, and Kwan-Liu Ma. 2018. Chart Constellations: Effective Chart Summarization for Collaborative and Multi-User Analyses. *Computer Graphics Forum* 37, 3 (2018), 75–86. https://doi.org/10.1111/cgf.13402

[93] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[94] Mark S Young, Neville A Stanton, and Don Harris. 2007. Driving automation: learning from aviation about design philosophies. *International Journal of Vehicle Design* 45, 3 (2007), 323–338.

[95] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[96] Jian Zhao, Michael Glueck, Simon Breslav, Fanny Chevalier, and Azam Khan. 2016. Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 261–270.

[97] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. 2017. Supporting Handoff in Asynchronous Collaborative Sensemaking Using Knowledge-Transfer Graphs. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 340–350.