Exploring The Role of Local and Global Explanations in Recommender Systems

Marissa Radensky*
radensky@cs.washington.edu
University of Washington

Doug Downey
dougd@allenai.org
Allen Institute for Artificial
Intelligence & Northwestern
University

Kyle Lo
kylel@allenai.org
Allen Institute for Artificial
Intelligence

Zoran Popović
zoran@cs.washington.edu
University of Washington

Daniel S. Weld
weld@cs.washington.edu
University of Washington & Allen
Institute for Artificial Intelligence

ABSTRACT

Explanations are well-known to improve recommender systems' transparency. These explanations may be local, explaining an individual recommendation, or global, explaining the recommender model in general. Despite their widespread use, there has been little investigation into the relative benefits of these two approaches. Do they provide the same benefits to users, or do they serve different purposes? We conducted a 30-participant exploratory study and a 30-participant controlled user study with a research-paper recommender system to analyze how providing participants local, global, or both explanations influences user understanding of system behavior. Our results provide evidence suggesting that both explanations are more helpful than either alone for explaining how to *improve* recommendations, yet both appeared less helpful than global alone for efficiency in *identifying* false positives and negatives. However, we note that the two explanation approaches may be better compared in the context of a higher-stakes or more opaque domain.

1 INTRODUCTION

Recommender systems are used daily by millions of people, and explanations that clarify a recommender's behavior are well-known to improve users' perceptions of the recommender's usefulness [2, 4, 5, 13, 14, 29, 54, 55, 57], controllability [2, 17, 29, 35], trustworthiness [1, 6, 14, 35, 37, 47], and transparency [2, 8, 17, 29, 33, 35, 37, 53]. Some recommenders provide users with *local* explanations describing why a specific item is recommended [10, 35]. Others give users a *global* explanation describing how recommendations are selected by the system overall [28, 44]. Still others show *both* explanations, which can be presented separately [1, 2, 27, 30, 40, 45, 54] or in a unified manner [4–6, 11, 17, 49].

Despite the widespread use of local and global explanations in recommender systems, to the best of our knowledge there has been no investigation into how each type of explanation influences the transparency of a recommender system. Recommenders often require feedback in order to provide high quality recommendations. Do the two explanation types play complementary roles in helping users understand how the system may improve recommendations?

Are local explanations used differently if global explanations are also present, or vice versa? Is one explanation type better for detecting false positive or false negative recommendations? We examine these questions and more using the recommender Semantic Sanity, which allows users to create recommendation feeds of computer-science research papers.

In summary, we make the following contributions:

- A formative study regarding how to present local and global explanations in a research-paper recommender.
- An exploratory study and controlled user study, each with 30 computer-science researchers, using Semantic Sanity to investigate several hypotheses surrounding three conditions: local, global, and local-plus-global explanations.
- Evidence suggesting that 1) both explanations help users explain how to improve recommendations better than either alone, but 2) both is less helpful than global alone for efficiency in identifying false positives and negatives. Also, 3) users prefer less diverse local explanations when a global explanation is also available.

2 RELATED WORK

2.1 Local and Global Explanations in Machine Learning

In machine learning broadly, global explanations explain how a model behaves generally, while local explanations explain a single model output, as first distinguished by Ribeiro et al. [47]. With respect to model transparency, local and global explanations have been studied from several perspectives. Some works find that local explanations have advantages over global explanations. Ribeiro et al. [47] established that local explanations more easily achieve model faithfulness. Similarly, Guidotti et al. [19] found that local explanations are more accurate and less complex than global explanations in simulating a black-box model's decisions. Other studies discuss benefits from both local and global explanations. For an image classification task, Mishra et al. [41] observed that local and global explanations both aid users in estimating model confidence and gauging their own confidence in the model output. Huber et al. [25] found that participants shown both local and global explanations instead of either alone performed best in evaluating

---

reinforcement-learning agents. For a task predicting risk of recidivism, one study demonstrated that local explanations are more helpful in discerning algorithmic fairness on a case-by-case basis, yet global explanations are perceived as more useful for understanding the model [12]. Another study showed that data scientists found both local and global explanations useful in trying to understand a model. However, novices preferred local explanations, while experts preferred global explanations [24]. In addition, Kopitar et al. [31] saw evidence that local interpretability provides additional insight over global interpretability in machine learning models for type 2 diabetes mellitus screening. We build on these works to address local and global explanations for transparency of recommender systems in particular. Recommender systems differ from most AI systems in that their output cannot be objectively evaluated as correct or not. Local and global explanations may be used differently when users must *subjectively* decide whether or not a recommendation is good and utilize that information to provide feedback to the recommender.

2.2 Explanations for Appropriate Trust of AI

When someone has appropriate trust in an intelligent system, they recognize when it is correct or not [36]. Studies have shown that appropriate trust of AI can be difficult to attain through explanations [3, 9, 21, 26, 41, 56, 58]. When the AI and user have similar decision-making performance, two studies found that, when compared to AI confidence, explanations do not improve team performance [3] or trust calibration [58] respectively. However, in a study surrounding a question-answering task, explanations did help users develop more appropriate trust in comparison to AI confidence [16]. The authors note this difference may be caused by the fact that, unlike other tasks, this one's explanations provide users with previously unseen information rather than just weighting of already seen evidence. Another study found that the timing and presentation of explanations can encourage more or less appropriate trust, but users prefer settings inducing less appropriate trust [7]. With regards to local and global explanations, Huber et al. [25] saw that both may help establish appropriate trust in reinforcement-learning agents. Meanwhile, Mishra et al. [41] found that, in an image classification task, global explanations were slightly less helpful for estimating model confidence in true positives compared to false positives. Here, we explore how local and global explanations may influence appropriate trust of a recommender by investigating if they help users to identify false positive and false negative recommendations.

2.3 Dimensions of AI Explanations

AI explanations have been designed and studied along several dimensions in addition to that of local and global explanations. For one, they may be generated using model-agnostic [38, 47] or model-specific [34, 38, 39] methods. They may be factual, explaining why a certain model outcome occurred, or contrastive, explaining why another outcome did not occur; they may also be counterfactual, explaining how another outcome could have occurred instead [18, 22, 51]. Furthermore, they have been investigated with regards to diverse user attributes such as their domain expertise [42, 48, 50], machine-learning expertise [24, 42, 48, 52], stakeholder group (e.g., developers, end users) [46], cognitive skills [40], and personality
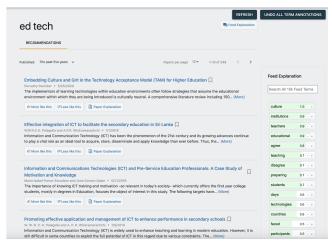

Figure 1: A paper recommendation in the local condition of Study 2, with the local explanation open at the bottom.

traits (e.g., openness, neuroticism) [32, 40]. AI explanations have additionally been compared in terms of various modalities such as visualization [16, 52], text [52], and audio [16]. Other studies have varied the length [16] and number [32] of explanations to observe the impact on cognitive load, as well as the toggle-ability and timing of explanations to reduce user biasing [7]. Two explanation dimensions that have been studied specifically in the context of recommender systems are style (e.g., social-based, content-based) [15, 32] and actionability [28, 30, 35, 43, 57]. In this paper, the local and global explanations are content-based and actionable.

3 STUDY 1: FORMATIVE STUDY FOR SYSTEM DESIGN

We first ran a formative study presenting design mockups for the recommender Semantic Sanity to six computer-science researchers in order to determine how best to present local, global, and local-plus-global explanations. These explanations are terms (unigrams and bigrams) from papers, a form of the common content-based explanation [1, 2, 4, 15, 27, 32]. The global terms are those with the most positive weights in the linear model for selecting paper recommendations. The local terms are those with the most positive product of model weight and TF-IDF value for the term's associated paper, and we use LIMEADE's approach [35] for introducing some randomness to diversify the local terms. We found a majority of participants preferred that local and global explanations be toggle-able and that they be presented in a unified manner when both available. Most participants also desired that they be actionable, meaning the user may directly manipulate the explanation widget to provide feedback to the recommender system [35]. Furthermore, participants easily understood that when local explanations had varying numbers of terms, only the most significant terms were shown, so we allowed variable-length local explanations. Within the constraint of two to four terms total, the system added terms to the local explanation until the term weights hit a plateau, meaning the explanation had the most salient terms.

Figure 2 shows the resulting interface for the local-plus-global condition. In all conditions, users can like or dislike papers and give feedback on terms considered by the model. In the **local-plus-global** condition, the "Feed Explanation" button at the top allows users to close or reopen a sidebar containing the global explanation. The sidebar presents the top 80 terms most related to the feed and allows users to search all 15,000 terms. Users can adjust terms' importance to the feed using the plus and minus buttons. The user may adjust terms' ratings between 0.0 and 1.0; one click adds or subtracts 0.1 to a term's rating. Additionally, users can click the "Paper Explanation" button under each paper to display a local explanation. This surfaces two to four paper-relevant terms at the top of the sidebar, and users can click the carrot underneath them to put the
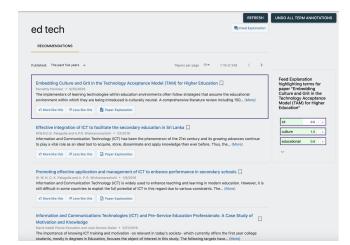
Figure 2: User interface for the local-plus-global condition in Study 2. Left: default layout. Right: layout when a local explanation is open. Irrespective of condition, the following features are present: "(More)" button under each paper to see its full abstract, "More like this"/"Less like this" buttons under each paper to provide feedback, a bookmark button next to each paper to save it, a "Refresh" button to apply user feedback, an "Undo Term Annotations Applied By Refresh" button shown directly after refreshing to undo all term annotations applied in the last refresh, and an "Undo All Term Annotations" button to return all terms to their original ratings.

Table 1: Metrics considered in Study 2, with corresponding hypotheses defined in Section 4.1.1. The questions are 7-point Likert-type questions. The two log file metrics (LFM) are from a click log file.

Hypo.	Metric ID	Metric
-	Q0: feed success	"The recommendation feed helps me find relevant papers."
H1	Q1: past actions	"The explanation(s) help me to understand why the system returned the papers it did."
H2	Q2: future actions	"The explanation(s) help me to anticipate what kinds of papers the system will return in the future."
H3	Q3: understand me	"The explanation(s) help me to know when the system doesn't understand my interests."
H4	Q4: change behavior	"When the feed is not completely relevant, I can explain how I would like the system to behave to be more relevant."
H5	Q5: false pos paper	"The explanation(s) help me to determine whether a **paper** is relevant or irrelevant."
	Q6: false pos term	"The explanation(s) help me to understand which **term** might cause an irrelevant paper to appear in my feed."
	LFM1	% of annotated terms that are annotated negatively
H6	Q7: false negative	"The explanation(s) help me to understand how likely the feed is to **miss papers** that I'd consider relevant."
H7	Q8: local diversity	"I would like the Paper Explanations to cover a less diverse set of terms, focusing more on the highest-rated terms."
H8	LFM2	# of annotated terms

terms in context of the feed's other top terms. The **global** condition looks similar but does not include the "Paper Explanation" buttons. In the **local** condition, users can click the "Paper Explanation" button under each paper to reveal or hide two to four terms explaining why the paper was recommended (Figure 1). They can also open all local explanations with a "View All Paper Explanations" button.

4 STUDY 2: EXPLORATORY STUDY

4.1 Study Design

4.1.1 Hypotheses. The objective of Study 2 was to explore how people use local and global explanations in a research-paper recommender system. The first six hypotheses concern transparency and are inspired by target purposes of AI explanations enumerated in previous work [20, 23]. These hypotheses state that there is at least one paired difference among the local, global, and local-plus-global conditions with regards to how helpful they are for... **H1**: understanding the recommender's past actions, **H2**: understanding the recommender's future actions, **H3**: knowing how well the

system understands the user, **H4**: understanding how the system can improve, **H5**: identifying false positives, and **H6**: identifying false negatives. The final two hypotheses address how users' interactions with the explanations are affected by the explanation types provided. **H7**: There is a difference between local and local-plus-global with regards to how diverse users want the local explanation terms to be, and **H8**: there is at least one paired difference among local, global, and local-plus-global with regards to how much feedback users provide on their explanations. The hypotheses' related 7-point Likert-type questions and log file metrics are outlined in Table 1.

4.1.2 Participants and Treatments. Thirty researchers who read at least one computer-science research paper each month interacted with the recommender system in a half-hour to one-hour session and were compensated with $25 Amazon gift cards. Fifteen participants went through both the global and local conditions in randomized order, and the other 15 interacted only with the local-plus-global condition. We did not include a baseline condition (no explanation) because the importance of explanations to recommender transparency is well-established [2, 8, 17, 29, 33, 35, 37, 53]. When signing up for the study, each participant provided two topics of interest, which would act as their feed topics.

4.1.3 Procedure. We first presented participants with a condition-specific slide tutorial. We then instructed participants to navigate to a specified link in order to access the recommender system. Clicks during the interaction with the system were recorded in a log file. Next, participants started their recommendation feed about their preset feed topic by selecting 4 seed papers, found using keyword search. Once they narrowed down their seed papers, they named and generated the feed. The participants' objective was to make the recommendation feed as relevant to them as possible. They had 15 minutes to do so, but if they felt that the feed was not going to become any more relevant before 15 minutes had passed, they stopped early. We also asked participants to think aloud as they interacted with the system in case there were any helpful insights into their interactions or they needed a reminder of how to use a certain system feature.

At the end of each condition, participants filled out a Google Forms survey without looking at the system. The survey first asked for short answers regarding in what situations, if any, the participant found each type of explanation useful. If the participants had any other thoughts on the explanations, they provided those as well. After, they answered the Likert-type questions discussed in Table 1. Lastly, participants returned to their feed and categorized the final top ten papers as relevant, neutral, or irrelevant. However, since this data depended heavily on factors other than successful feed curation (e.g. the number of papers published on the feed topic), we did not utilize it.

4.2 Results and Discussion

4.2.1 Quantitative Results and Discussion. We organize our discussion of quantitative results around the hypotheses and metrics in Table 1. For the Likert-type questions, we compared the local and global conditions using the within-subjects two-tailed Wilcoxon signed-rank test and the remaining condition pairs using

the between-subjects two-tailed Mann-Whitney-Wilcoxon test. Violin plots for these results are presented in Figure 3. For the log file metrics, we analyzed all condition pairs using a one-way ANOVA test. The significance threshold was p < 0.05. Though all the results were insignificant after Bonferroni corrections, results for **H4** and **H7** would be significant without corrections.

Regarding **H4**, participants in the local-plus-global condition demonstrated more confidence than the participants in the global (p=0.015, uncorrected) or local (p=0.030, uncorrected) condition in explaining how they would like the system to behave to be more relevant. However, there was no difference indicated between local and global. Thus, the results suggest that **local and global explanations together are better than either alone for helping users understand how the recommender system can improve**. While similar results have been shown in other machine learning systems [24, 25], this is a distinct insight for recommender systems because, unlike those other systems, recommenders do not have objectively correct or incorrect output. The recommender's output is judged and rated according to the user's own standards. This personal form of judgment may benefit more or less from local and global explanations.

In order to create appropriately transparent interactions, a designer needs to know what kinds of information users seek from local explanations. The result for **H7** suggests that the ideal content of local explanations depends on whether or not a global explanation is present. In particular, **participants desired *less diverse* and more consistent local explanations when the global explanation was also present** (p=0.038, uncorrected). This may be a consequence of the "explanation-action trade off" [35], which refers to how actionable local explanations in recommender systems must balance two competing goals: 1) returning the most accurate explanations and 2) affording more opportunities for users to adjust the model. The goals are at odds because the most accurate local explanations often share the same terms and thus provide fewer chances for users to adjust the model. We address this in Semantic Sanity by explicitly introducing randomness to diversify the local explanations, as Lee et al. does [35]. When local explanations are alone, they are the only means by which users can act on the system, so greater diversity is appreciated by users. In contrast, when an actionable global explanation is also present, users no longer depend on local explanations for adjusting the model and can use them more as a means of explanation, which users may expect to be consistent with the global explanation and thus less random.

4.2.2 Qualitative Results and Discussion. In their short-answer responses, **participants commented more often that they forgot or did not find much use for the local explanations as compared to the global explanation**. Of the 30 participants, 9 mentioned either forgetting local explanations or using them rarely, whereas only one participant mentioned not using the global explanation. However, this difference may be due in part to a user interface design issue, which is described in Section 5.1.2.

Participants also noted that local and global explanations may serve different purposes in terms of research exploration and discovery. Four participants explained that **the ability to adjust the importance of the global explanation terms was useful to help**
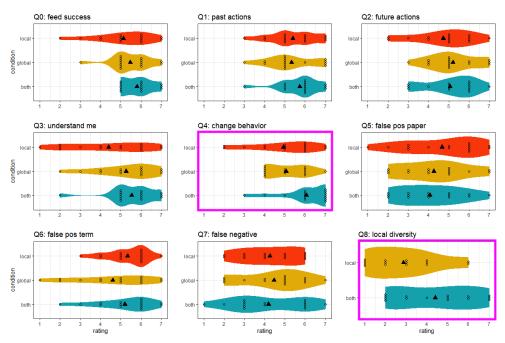
Figure 3: Study 2 results for each Likert-type question and condition. 1 indicates "strongly disagree," while 7 indicates "strongly agree." Each triangle represents the mean response for the given question and condition, while the circles within each plot represent individual responses. Q4: With both explanations rather than only global (p=0.015, uncorrected, two-tailed Mann-Whitney-Wilcoxon) or only local (p=0.030, uncorrected, two-tailed Mann-Whitney-Wilcoxon), participants were more confident in explaining how they would like the system to behave to be more relevant. Q8: Participants desired less diverse local explanations when the global explanation was also present (p=0.038, uncorrected, two-tailed Mann-Whitney-Wilcoxon).

them avoid unintended bias towards specific authors or topics. For example, P17 noted, "*The system seemed to be suggesting a particular author and listed that in the feed explanation column. I reduced that so that I could have a more unbiased feed of people I don't often read....*" Two participants also mentioned that **the global explanation allowed them to introspect about their own research interests**. For instance, P11 commented, "*[Global] gave me a better idea of what my inputs... seemed to have in common.*" On the other hand, two participants found **local explanations were useful for characterizing unexpected interesting papers**. P24 wrote, "*There was a paper suggested to me that I found relevant, but I was also surprised to find it in my recommendation list... [Local] was useful for me to check out why that paper was recommended (so that I can see more such papers!).*"

5 STUDY 3: CONTROLLED USER STUDY

5.1 Study Design

5.1.1 Hypotheses. Study 2 provided suggestive evidence that both explanations are better than either alone for understanding how the recommender may improve. Study 3's objectives were to confirm this point and to investigate *how* local and global explanations complement one another to help users understand recommender output. Thus, Study 3's hypotheses were as follows. **H9**: Local is better than global for identifying false positive recommendations, **H10**: global is better than local for identifying false negative recommendations, and **H11**: both are better than either alone for understanding how

the recommender may improve. **H11** exists to confirm the suggestive results from Study 2. **H9** and **H10** reflect a framework for how local and global may complement each other to make the recommender more transparent. The hypotheses' associated metrics are provided in Table 2 and are described further in Section 5.2.

5.1.2 Participants and Treatments. In the same manner as in Study 2, thirty computer-science researchers were recruited and separated into treatments. Minimal changes were made to the design of the explanations. Their titles were updated to be purple for emphasis, and they were renamed to better draw users' attention. The local explanations were renamed from "Paper Explanation" to "Why This Paper," and the global explanation was renamed from "Feed Explanation" to "Why This Feed." Also, as is described in Section 5.1.3, Study 3's procedure no longer required participants to curate recommendation feeds, so the only clickable buttons were for looking at the explanations and paper abstracts. The remaining buttons were still included to provide participants with context for how the recommender system would work overall.

Furthermore, the local-plus-global condition was updated so that local and global explanations were presented separately. This update was made because, in Study 2, the unified presentation of local and global may have led participants to focus less on local explanations. In the local-plus-global condition, the local explanations could only be opened one-by-one. On the other hand, in the local condition, participants could open as many explanations as they wanted. Perhaps due to this user-interface design, even though

Table 2: Metrics considered in Study 3, with corresponding hypotheses defined in Section 5.1.1. The question is a 7-point Likert-type question. The score calculations are described in Section 5.2.

Hypo.	Metric ID	Metric
H9	M1	score on false-positive survey (between -42 and 42)
H10	M2	score on false-negative survey (0 or 1)
H11	Q9	"I can explain how the system should be updated to be more relevant."

participants in the local condition could open all local explanations with a single click, they still opened an individual local explanation 9.3 times on average, while participants in the local-plus-global condition opened an individual local explanation only 2.7 times on average.

When the participants were asked to choose topics of interest for their feeds in Study 2, the feed topics varied largely in breadth and familiarity. This may have hindered our ability to observe significant results in Study 2. As a result, Study 3's participants were randomly assigned to one of two preset feeds for each condition: "misinformation on social media" or "educational technologies for demographically diverse users." These feed topics were chosen based on three criteria. First, in order for participants from varying research areas to engage with the feed, the topic and its explanations needed to use limited jargon. Second, the topic needed to be specific enough that false positives occurred within the top twenty papers of the feed. Third, the topic needed to be broad enough so that a cluster of false negatives emerged. For example, in the "misinformation on social media" feed, true positives were exclusively about *coronavirus-related* misinformation, so any papers discussing misinformation on social media not related to coronavirus formed a cluster of false negatives. The preset feeds were seeded with five papers selected so that the feeds would fit the criteria just mentioned.

Three annotators classified the top 20 papers of each 250-paper feed as false or true positives and the bottom 50 papers of each feed as false or true negatives, based on the papers' titles and abstracts. Only papers upon which there was unanimous agreement were added to the pool of papers that the participants could encounter. The original local explanations for each annotated paper were then kept constant, so that no new randomized terms were introduced for diversification.

Subsequently, the twenty-first paper from the "educational technologies for demographically diverse users" feed was added to the pool of papers in order to have enough true positive papers for the study. Also, the "misinformation on social media" feed had ten false negatives. Two did not belong to the cluster consisting of papers discussing *non-coronavirus* misinformation on social media. To make sure all participants interacting with this feed would see a false negative from the same cluster, these two false negatives were removed from the pool of papers participants could see.

5.1.3 Procedure. Participants first opened a link to the recommender system. For each condition, they then logged into one of two accounts to access a preset feed with six recommendations. Next, we gave them a condition-specific tutorial on using the system. The participants then answered three Google-Forms surveys to address each hypothesis.

H9 was addressed first with a false-positive survey. The survey asked participants to label each of the six paper recommendations in the feed as relevant or not and rate how confident they were in their answers on a 7-point scale. The recommendations were randomly ordered and selected such that half would be false positives. About half of all the true positives had optimal local explanations containing information pertinent to both aspects of the given feed topic. For instance, in the "misinformation on social media" feed, the optimal local explanation may have the term "fake news" related to "misinformation" as well as the term "twitter" related to "social media." To make sure this category of true positive was represented accordingly, one of these true positives was randomly chosen to be included in each participant's feed.

H10 was addressed next with a false-negative survey. The survey presented participants with three new paper recommendations for the feed. Two were true positives and one was a false negative. The survey asked participants to rank these papers based on how they believed the recommender system *would rather than should* rank the papers. Ideally, the participant would be able to recognize that the false negative paper would be ranked last by the system.

Finally, **H11** was addressed with a survey asking participants to answer the 7-point Likert-type question **Q9**. The survey also asked participants to explain to a software developer how to make the recommendations more relevant, but we found that participants did not understand this question as intended, so it was discarded.

5.2 Results and Discussion

We organize our discussion of results around the hypotheses and metrics discussed in Table 2. The false-positive survey score **M1** was calculated as follows. For each of the six recommendations, if the participant classified it correctly as relevant or not to the feed topic, 1 multiplied by their confidence (1 to 7) was added to their cumulative score. If the participant classified it incorrectly, -1 multiplied by their confidence was added. The false-negative survey score **M2** was 1 if the false negative paper was ranked below the two true-positive papers and 0 if not. For **Q9**, we compared the local and global conditions using the within-subjects two-tailed Wilcoxon signed-rank test and the other condition pairs using the between-subjects two-tailed Mann-Whitney-Wilcoxon test. For **M1** and **M2**, we analyzed all condition pairs using a one-way ANOVA test. The significance threshold was $p < 0.05$. All results were insignificant.

Regarding **H11**, the slight change in wording of the Likert-type question **Q9** in Study 3 as compared to **Q4** in Study 2 may have implied that, in order to respond affirmatively, the participant needed a more technical rather than merely conceptual understanding of how the system could improve its recommendations. Also, participants may have had more trouble conceptualizing how the system

should improve when they did not choose the feed topic. Both of these points would explain the overall lower average response to **Q9** of 4.83, as compared to the average response to **Q4** of 5.36.

Regarding **H9** and **H10**, while we did not find any significant differences among the conditions with respect to how well participants identified false positives or negatives, we did observe uncorrected significant differences among the conditions in terms of how *quickly* participants completed the false-positive and false-negative surveys for the "misinformation on social media" feed, as shown in Figure 4. Each participant's time spent on each survey was rounded to the nearest half-minute. These results are uncorrected because they were not pre-registered for analysis. Twenty-two participants completed a condition using the "misinformation" feed (8 in the global condition, 7 in each of the other conditions). We analyzed all condition pairs for the feed using a one-way ANOVA test followed by a Tukey HSD test.

With regards to the "misinformation" feed's false-positive survey, participants in the local-plus-global condition completed the survey slower than those in the global (p=0.020, uncorrected) and local (p=0.045, uncorrected) conditions. These results suggest that **providing both explanations rather than either alone causes users to identify false positives more slowly**. This may simply be due to the fact that there is more information for users to consider when both explanations are available.

With regards to the "misinformation" feed's false-negative survey, participants in the global condition completed it faster than the participants in the local-plus-global condition (p=0.018, uncorrected). Though insignificant, participants in the global condition also completed the survey faster than participants in the local condition (p=0.135, uncorrected). The first result suggests that **providing only a global explanation rather than both explanations helps users identify false negatives more quickly**. This makes sense for two reasons. Firstly, when both explanations are present, there is more information for users to evaluate. Secondly, in comparison to the local explanations, the global explanation's top terms provide users a straightforward indication of which terms the model may be considering too important or unimportant, which can cause false negatives. With only local explanations, users must estimate which terms are most important to the feed by comparing several local explanations.

There are a few possible reasons why we did not observe the same results for the "educational technologies for demographically diverse users" feed. For one, participants generally noted that this topic was more difficult to understand. With respect to the false negative finding, this feed's false negative cluster was the result of an *over-specification* rather than an *irrelevant specification*. The cluster consisted of papers related to educational technologies for *gender-diverse* users. In the global explanation, the only top terms related to diversity were related to ethnic rather than gender diversity. This issue is likely more difficult to note because, unlike the irrelevant specification for "covid" in the "misinformation" feed's global explanation, the term "cultural" in this feed *does* contribute to the feed topic. Terms like "cultural" merely cause the model to overfit to ethnic diversity when terms related to gender diversity should also be included. Thus, global explanations may only help users identify false negatives more quickly when they are the result of an irrelevant specification, as opposed to an over-specification.
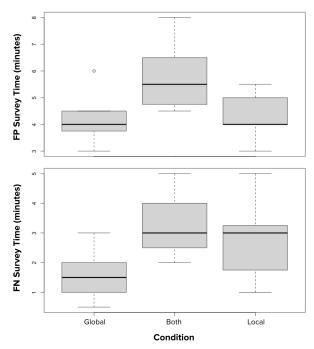

Figure 4: **How much time participants spent on the false-positive (top) and false-negative (bottom) surveys as a function of the explanation condition, under the "misinformation on social media" feed. Top: Participants spent more time on the false-positive survey when both explanations were present as compared to only global (p=0.020, uncorrected, one-way ANOVA) or only local (p=0.045, uncorrected, one-way ANOVA). Bottom: Participants spent less time on the false-negative survey when only global was present as compared to both (p=0.018, uncorrected, one-way ANOVA) or only local (p=0.135, uncorrected, one-way ANOVA). Providing the global explanation alone thus appears more helpful than providing both explanations for identifying false positives and negatives efficiently.**

Lastly, this feed's over-specification in the global explanation was less obvious with fewer related terms than the "misinformation" feed's irrelevant specification.

However, in a follow-up formative study that introduced time constraints for completing the false-positive and false-negative surveys, participants were not evidently better at identifying false positives or negatives in one explanation setting versus another. There are a couple reasons why this could be. For one, computer-science researchers are already accustomed to evaluating the relevance of paper recommendations without explanations, and perhaps often based on titles alone. Explanations may not be useful for identifying false positives and negatives when the recommendations are already sufficiently transparent, especially in a generally lower-stakes situation like browsing research papers. In addition, participants were not necessarily invested in or familiar with the feed topics, as they were not selected by them. As noted in Study 2, there is a trade-off in studying personalized feed topics because they can vary

in breadth and familiarity. Nonetheless, given the lack of meaningful results in Study 3 and the inherent individualized nature of recommenders, having participants engage with personal recommendations seems essential to studying recommenders. In a similar vein, since the feed topics were chosen to be accessible to all computer-science researchers, identifying false positives and negatives may have been uncommonly easy.

6 CONCLUSION, LIMITATIONS, AND FUTURE WORK

Following a formative study to determine how content-based local and global explanations should be presented in a research-paper recommender system, we conducted an exploratory study comparing the use of the two explanation approaches in this system. We found evidence suggesting that each explanation type plays a unique role in augmenting the system's transparency and influences how the other is used for understanding the system. Specifically, our results suggest that

- Providing both explanations rather than either alone ensures users reach the best understanding of how the recommender can improve, and
- Users prefer more diverse local explanations when they are presented alone compared to when a global explanation is also available.

The study also provided qualitative evidence that, in the domain of research papers, local and global explanations may be useful for a purpose other than determining recommendation relevance-exploration and discovery of research.

In a subsequent controlled user study, we investigated *how* local and global explanations may complement one another to help users understand their recommendations, in particular by revealing false positives and false negatives. While we did not find any significant differences between the two explanations in terms of utility in identifying false positives or negatives, we did observe evidence suggesting that

- Providing both explanations rather than either alone slows users' identification of false positives, and
- Providing a global explanation alone rather than both explanations quickens users' identification of false negatives caused by unnecessary specifications.

However, a follow-up formative study did not corroborate these findings.

Limitations of this work include that 1) the user studies were small-scale and 2) only one recommendation domain (research papers) and explanation style (content-based) were studied. Future work may study the use of local and global explanations for more opaque recommendations such as author or artist recommendations; an explanation is less necessary when the recommendation itself summarizes its contents, as is the case with paper recommendations. Future research may also explore how these explanations are used in higher-stakes recommendation settings, such as education or healthcare, in which explanations likely bear greater importance. Finally, future work may investigate how local and global explanations are used for purposes other than clarifying recommendation relevance, such as discovery of more diverse recommendations.

REFERENCES

[1] J. Ahn, P. Brusilovsky, J. Grady, D. He, and Sue Yeon Syn. 2007. Open user profiles for adaptive news systems: help or harm?. In *WWW '07*.

[2] F. Bakalov, M. Meurs, B. König-Ries, Bahar Sateli, R. Witte, G. Butler, and A. Tsang. 2013. An approach to controlling user models and personalization effects in recommender systems. In *IUI '13*.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. 2020. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779* (2020).

[4] Svetlin Bostandjiev, J. O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *RecSys '12*.

[5] Svetlin Bostandjiev, J. O'Donovan, and Tobias Höllerer. 2013. LinkedVis: exploring social and semantic career recommendations. In *IUI '13*.

[6] S. Bruns, André Calero Valdez, Christoph Greven, M. Ziefle, and U. Schroeder. 2015. What Should I Read Next? A Personalized Visual Publication Recommender System. In *HCI*.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[8] Joseph Chee Chang, Nathan Hahn, Adam Perer, and A. Kittur. 2019. SearchLens: composing and capturing complex user interests for exploratory search. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[9] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[10] Henriette Cramer, V. Evers, Satyan Ramlal, M. V. Someren, L. Rutledge, N. Stash, Lora Aroyo, and B. Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18 (2008), 455–496.

[11] Laura Devendorf, J. O'Donovan, and Tobias Höllerer. 2012. TopicLens : An Interactive Recommender System based on Topical and Social Connections.

[12] J. Dodge, Q. Liao, Y. Zhang, R. Bellamy, and C. Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[13] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and D. Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[14] A. Felfernig and B. Gula. 2006. An Empirical Study on Consumer Behavior in the Interaction with Knowledge-based Recommender Applications. *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)* (2006), 37–37.

[15] G. Friedrich and M. Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Mag.* 32 (2011), 90–98.

[16] Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human Evaluation of Spoken vs. Visual Explanations for Open-Domain QA. *arXiv preprint arXiv:2012.15075* (2020).

[17] Brynjar Gretarsson, J. O'Donovan, Svetlin Bostandjiev, C. Hall, and Tobias Höllerer. 2010. SmallWorlds: Visualizing Social Recommendations. *Computer Graphics Forum* 29 (2010).

[18] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[19] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *ArXiv* abs/1805.10820 (2018).

[20] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017), 2.

[21] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *arXiv preprint arXiv:2005.01831* (2020).

[22] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809* (2018).

[23] R. Hoffman, S. Mueller, G. Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *ArXiv* abs/1812.04608 (2018).

[24] Fred Hohman, Andrew Head, R. Caruana, Robert DeLine, and S. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[25] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2020. Local and global explanations of agent behavior: integrating strategy summaries with saliency maps. *arXiv preprint arXiv:2005.08874* (2020).

[26] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[27] Y. Jin, N. Tintarev, and K. Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. *Proceedings of the 12th ACM Conference on Recommender Systems* (2018).

[28] Antti Kangasrääsiö, D. Glowacka, and Samuel Kaski. 2015. Improving Controllability and Predictability of Interactive Recommendation Interfaces for Exploratory Search. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (2015).

[29] Bart P. Knijnenburg, Svetlin Bostandjiev, J. O'Donovan, and A. Kobsa. 2012. Inspectability and control in social recommenders. In *RecSys '12*.

[30] Bart P. Knijnenburg, Niels J. M. Reijmer, and M. C. Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *RecSys '11*.

[31] Leon Kopitar, Leona Cilar, Primoz Kocbek, and Gregor Stiglic. 2019. Local vs. Global Interpretability of Machine Learning Models in Type 2 Diabetes Mellitus Screening. In *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*. Springer, 108–119.

[32] Pigi Kouki, James Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. 2019. Personalized explanations for hybrid recommender systems. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[33] T. Kulesza, S. Stumpf, M. Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *CHI '12*.

[34] Will Landecker, Michael D Thomure, Luís MA Bettencourt, Melanie Mitchell, Garrett T Kenyon, and Steven P Brumby. 2013. Interpreting individual classifications of hierarchical networks. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 32–38.

[35] B. Lee, Kyle Lo, Doug Downey, and Daniel S. Weld. 2020. Explanation-Based Tuning of Opaque Machine Learners with Application to Paper Recommendation. *ArXiv* abs/2003.04315 (2020).

[36] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[37] Tianyi Li, Gregorio Convertino, Ranjeet Kumar Tayi, and Shima Kazerooni. 2019. What data should I protect?: recommender and planning support for data security analysts. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[38] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[39] David Martens, Johan Huysmans, Rudy Setiono, Jan Vanthienen, and Bart Baesens. 2008. Rule extraction from support vector machines: an overview of issues and application in credit scoring. *Rule extraction from support vector machines* (2008), 33–63.

[40] Martijn Millecamp, Nyi Nyi Htun, C. Conati, and K. Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[41] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[42] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839* 1 (2018).

[43] Dario De Nart, F. Ferrara, and C. Tasso. 2013. Personalized Access to Scientific Publications: from Recommendation to Explanation. In *UMAP*.

[44] J. O'Donovan, B. Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *CHI*.

[45] Denis Parra and P. Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *Int. J. Hum. Comput. Stud.* 78 (2015), 43–67.

[46] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[47] Marco Tulio Ribeiro, Sameer Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[48] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *IUI Workshops*.

[49] James Schaffer, Tobias Höllerer, and J. O'Donovan. 2015. Hypothetical Recommendation: A Study of Interactive Profile Manipulation Behavior for Recommender Systems. In *FLAIRS Conference*.

[50] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[51] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access* 9 (2021), 11974–12001.

[52] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[53] Chun-Hua Tsai and P. Brusilovsky. 2017. Providing Control and Transparency in a Social Recommender System for Academic Conferences. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (2017).

[54] Chun-Hua Tsai and P. Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

[55] Chun-Hua Tsai and Peter Brusilovsky. 2020. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* (2020), 1–37.

[56] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[57] J. Vig, S. Sen, and J. Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.* 2 (2012), 13:1–13:44.

[58] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 295–305.