Exploring the Role of Local and Global Explanations in Recommender Systems

Marissa Radensky*
radensky@cs.washington.edu
University of Washington
Seattle, WA, USA

Doug Downey
dougd@allenai.org
Allen Institute for AI & Northwestern University
Seattle, WA, USA

Kyle Lo
kylel@allenai.org
Allen Institute for AI
Seattle, WA, USA

Zoran Popović
zoran@cs.washington.edu
University of Washington
Seattle, WA, USA

Daniel S. Weld
weld@cs.washington.edu
University of Washington & Allen Institute for AI
Seattle, WA, USA

ABSTRACT

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

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Computing methodologies → Machine learning**; • **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

explainable AI, human-AI interaction

ACM Reference Format:
Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popović, and Daniel S. Weld. 2022. Exploring the Role of Local and Global Explanations in Recommender Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3519795

1 INTRODUCTION AND RELATED WORK

Recommender systems are used daily by millions of people, and explanations that clarify their behavior are well-known to improve users' perceptions of their usefulness [2–4, 10, 11, 21, 35–37], controllability [2, 13, 21, 26], trustworthiness [1, 5, 11, 26, 27, 32], and transparency [2, 6, 13, 21, 25–27, 34]. Recommenders may provide users with *local* explanations describing why a specific item is recommended [7, 26], a *global* explanation describing how recommendations are selected overall [20, 30], or *both*, presented separately [1, 2, 19, 22, 28, 31, 35] or in a unified manner [3–5, 8, 13, 33].

Despite widespread use of local and global explanations in recommender systems, to the best of our knowledge there has been no investigation into how each influences recommender transparency. 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. [32]. Some works find that local explanations are more useful for model faithfulness [14, 32]. Others discuss benefits from both explanations in terms of understanding and evaluating models [9, 17, 18, 23, 29]. We build on these works to address how local and global explanations affect *recommender* transparency.

Recommenders 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 recommendations' relevance and provide feedback. Do the two explanation types play complementary roles in helping users understand how the system may improve? Is one better for detecting false positive or false negative recommendations? Are local explanations used differently if a global explanation is also present, or vice versa? We examine these questions and more using Semantic Sanity, a system that 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 system.
- An exploratory study and controlled user study, each with 30 computer-science researchers, using the recommender to investigate several hypotheses surrounding three conditions: local, global, and both explanations.

[*] Work done during AI2 internship and Ph.D. at University of Washington.

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3519795

- 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 efficiently identifying false positive and negative recommendations. Also, 3) users prefer less diverse local explanations when a global explanation is also available.

2 STUDY 1: FORMATIVE STUDY FOR SYSTEM DESIGN

We ran a formative study presenting design mockups to six computer-science researchers to determine how best to present local and global explanations in the research-paper recommender Semantic Sanity. These explanations are terms (unigrams and bigrams) from papers, a form of the common content-based explanation [1–3, 12, 19, 24]. Presenting the explanations as simple key terms, as in some other explainable recommenders [1, 3, 19], prevents them from adding too much clutter to the already information-heavy user interface. The global terms have the most positive weights in the recommender's linear model. The local terms have the most positive products of model weight and TF-IDF value for the associated paper; we use LIMEADE's approach [26] for introducing some randomness to diversify the local terms. A majority of participants preferred that local and global explanations be toggle-able, presented in a unified manner when both available, and actionable, meaning the user may manipulate the explanation widget to provide feedback [26]. Also, participants easily understood that when local explanations had varying numbers of terms, only the most significant terms were shown. Thus, within the constraint of two to four terms total, the system added terms to a local explanation until the term weights hit a plateau so that the explanation had the most salient terms.

Figure 1 shows the resulting interface for the local-global condition. In all conditions, users can like or dislike papers and give feedback on explanation terms. In the **local-global (LG)** condition, the "Feed Explanation" button opens a sidebar (open by default) containing the global explanation. The sidebar shows the top 80 feed terms and allows users to search all 15,000 terms. Users can adjust term ratings between 0.0 and 1.0 by using the plus and minus buttons to add or subtract 0.1. The "Paper Explanation" button under each paper displays a local explanation. This surfaces two to four paper-relevant terms at the top of the sidebar, and clicking the carrot underneath them puts the terms in context of the global explanation. The **global (G)** condition looks similar but does not include the "Paper Explanation" buttons. In the **local (L)** condition, the "Paper Explanation" button under each paper reveals two to four terms explaining why the paper was recommended (Figure 2), and a "View All Paper Explanations" button opens all local explanations.

3 STUDY 2: EXPLORATORY STUDY

3.1 Study Design

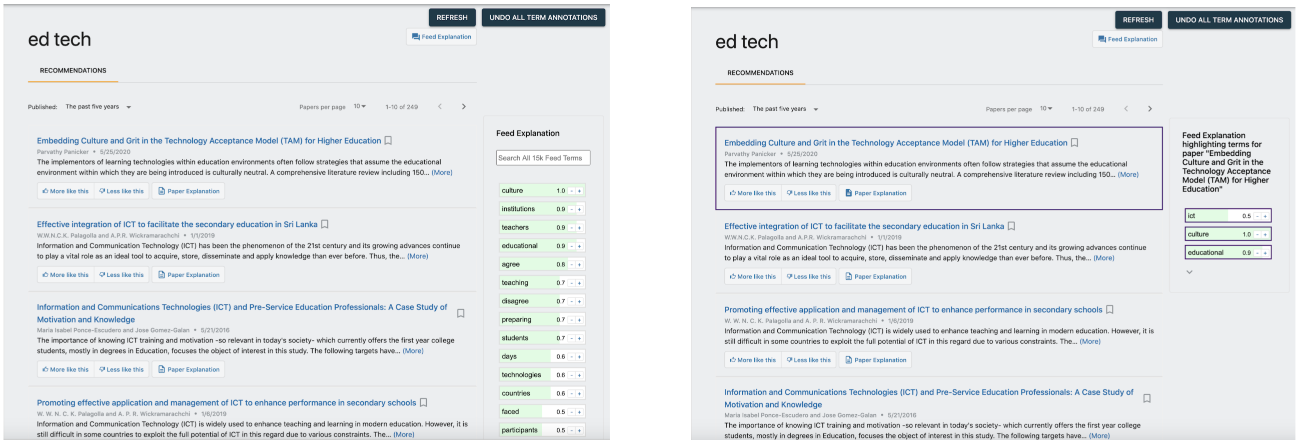
3.1.1 Hypotheses. Study 2's overarching research question **RQ1** was: how do users utilize local and global explanations in a research-paper recommender? The first six hypotheses relate to how the explanations affect the recommender's transparency and are inspired by target purposes of AI explanations enumerated in previous work [15, 16]. They state that there is at least one paired difference among the L, G, and LG conditions in terms of utility in... **H1**: understanding past recommender actions, **H2**: understanding future recommender actions, **H3**: understanding how well the system understands the user, **H4**: understanding how the system can improve, **H5**: identifying false positive recommendations, and **H6**: identifying false negative recommendations. The final two hypotheses address how users' interactions with the explanations are affected by the explanation types provided. **H7**: There is a difference between the L and LG conditions with regards to desired diversity of local explanation terms, and **H8**: there is at least one paired difference among the three conditions with regards to amount of explanation feedback. The hypotheses' metrics are in Table 1.

3.1.2 Participants and Treatments. Thirty researchers who read at least one computer-science research paper monthly interacted with the recommender for a half-hour to one-hour and were compensated with $25 Amazon gift cards. Fifteen participants received the G and L conditions in randomized order; the other 15 interacted only with the LG condition. There was no baseline condition (no explanation) because explanations' importance to recommender transparency is well-established [2, 6, 13, 21, 25–27, 34]. In signing up, participants provided two topics of interest for their feed topics.

3.1.3 Procedure. We first presented participants with a condition-specific slide tutorial and then instructed them to navigate to a specified link to access the recommender. Clicks were recorded in a log file. Participants started the recommendation feed about their chosen topic with 4 seed papers, found using keyword search, and named and generated the feed. The participants' objective was to make the feed as relevant to them as possible. They had 15 minutes to do so, but if they felt that it was not going to become any more relevant before 15 minutes passed, they stopped early. The number of seed papers and time limit for each feed were chosen so that the study would not be too long and fatiguing. 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 asked for short answers regarding in what situations, if any, they found each explanation type useful. It also asked for any other thoughts on the explanations. Next, they answered the Likert-type questions in Table 1. Lastly, they returned to the feed and categorized the final top ten papers as relevant, neutral, or irrelevant. However, this data depended heavily on factors other than successful feed curation (e.g. the number of papers published on the feed topic), so we did not utilize it.

3.2 Results and Discussion

3.2.1 Quantitative Results. Here we discuss results for **RQ1**'s related hypotheses and metrics (Table 1). For Likert-type questions, we compared the L and G conditions with the within-subjects two-tailed Wilcoxon signed-rank test and the remaining condition pairs with the between-subjects two-tailed Mann-Whitney-Wilcoxon test (Figure 3). For log file metrics, we analyzed all condition pairs with a one-way ANOVA test. The significance threshold was $p < 0.05$. Though all results were insignificant after Bonferroni corrections, results for **H4** and **H7** would be significant otherwise.



Feed Explanation

Embedding Culture and Grit in the Technology Acceptance Model (TAM) for Higher Education
Parvathy Panicker • 5/25/2020
The implementors of learning technologies within education environments often follow strategies that assume the educational environment within which they are being introduced is culturally neutral. A comprehensive literature review including 150... (More)

Effective integration of ICT to facilitate the secondary education in Sri Lanka
W.W.N.C.K. Palagolla and A.P.R. Wickramarachchi • 1/1/2019
Information and Communication Technology (ICT) has been the phenomenon of the 21st century and its growing advances continue to play a vital role as an ideal tool to acquire, store, disseminate and apply knowledge than ever before. Thus, the... (More)

Information and Communications Technologies (ICT) and Pre-Service Education Professionals: A Case Study of Motivation and Knowledge
Maria Isabel Ponce-Escudero and Jose Gomez-Galan • 5/21/2016
The importance of knowing ICT training and motivation -so relevant in today's society- which currently offers the first year college students, mostly in degrees in Education, focuses the object of interest in this study. The following targets have... (More)

Promoting effective application and management of ICT to enhance performance in secondary schools
W. W. N. C. K. Palagolla and A. P. R. Wickramarachchi • 1/6/2019
Information and Communication Technology (ICT) is widely used to enhance teaching and learning in modern education. However, it is still difficult in some countries to exploit the full potential of ICT in this regard due to various constraints. The... (More)

culture	1.0	-	+
institutions	0.9	-	+
teachers	0.9	-	+
educational	0.9	-	+
agree	0.8	-	+
teaching	0.7	-	+
disagree	0.7	-	+
preparing	0.7	-	+
students	0.7	-	+
days	0.6	-	+
technologies	0.6	-	+
countries	0.6	-	+
faced	0.5	-	+
participants	0.5	-	+

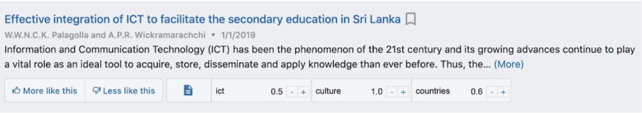
Feed Explanation highlighting terms for paper "Embedding Culture and Grit in the Technology Acceptance Model (TAM) for Higher Education"

ict	0.5	-	+
culture	1.0	-	+
educational	0.9	-	+

Figure 1: UI for Study 2 LG condition. Left: default layout. Right: layout when a local explanation is open. Irrespective of condition, these features are present: "(More)" buttons to see full abstracts, "More like this"/"Less like this" buttons for paper feedback, bookmark buttons to save papers, "Refresh" button to apply feedback, "Undo Term Annotations Applied By Refresh" button shown directly after refresh to undo all term annotations applied by refresh, and "Undo All Term Annotations" button to return all terms to original ratings.

Table 1: Study 2 metrics for hypotheses in Section 3.1.1. Questions are 7-point Likert-type questions. LFM means log file metric.

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



Effective integration of ICT to facilitate the secondary education in Sri Lanka
W.W.N.C.K. Palagolla and A.P.R. Wickramarachchi • 1/1/2019
Information and Communication Technology (ICT) has been the phenomenon of the 21st century and its growing advances continue to play a vital role as an ideal tool to acquire, store, disseminate and apply knowledge than ever before. Thus, the... (More)

ict	0.5	-	+
culture	1.0	-	+
countries	0.6	-	+

Figure 2: A paper recommendation in Study 2’s L condition, with the local explanation open at the bottom.

Regarding **H4**, participants in the LG condition demonstrated more confidence than participants in the G (W=55.5, p=0.015, uncorrected) or L (W=61.5, 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 the L and G conditions. This suggests that **both explanations together are better than either alone for helping users understand how the recommender can improve**. While similar results have been shown in other machine learning systems [17, 18], this is a distinct insight for recommenders, as their output is not objectively correct or not. Judging and rating output according to their own standards, users may benefit differently from local and global explanations.

To create appropriately transparent interactions, a designer needs to know what kinds of information users seek from local explanations. **H7**’s result suggests that local explanations’ ideal content 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** (W=62.5, p=0.038, uncorrected). This may be a consequence of the "explanation-action trade off" [26], which refers to how actionable local explanations without a global explanation in recommenders must balance two competing goals: 1) returning the most accurate (and often consequently more uniform) explanations and 2) affording more opportunities for users to adjust the model. We address this in Semantic Sanity by explicitly introducing randomness to diversify the local explanations, as in Lee et al. [26].

3.2.2 Qualitative Results. 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, nine mentioned either forgetting local explanations or using them rarely, whereas only one mentioned not using the global explanation. However, this difference may be due in part to a user interface design issue described in Section 4.1.2. Participants also noted that local and global explanations may serve different purposes in terms of research exploration. Four participants explained that **the ability to adjust the importance of the global explanation terms was useful to avoid unintended bias toward specific authors or topics**. 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 mentioned that **the global explanation allowed them to introspect about their own research interests**. P11 commented, "*[Global] gave me a better idea of what my inputs... seemed to have in common.*" On the other hand, two participants found that **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....*"

4 STUDY 3: CONTROLLED USER STUDY

4.1 Study Design

4.1.1 Hypotheses. Study 3's first research question aimed to reaffirm Study 2's main result. **RQ2**: are both explanations more helpful than either alone for explaining how to improve recommendations? To address **RQ2**, we have the hypothesis **H11**: both are better than either alone for understanding how the recommender may improve. Study 3's second research question sought to expand on Study 2's findings. **RQ3**: how do local and global explanations *complement* one another to help users understand recommender output? To address **RQ3**, we have two hypotheses reflecting a framework for how the two may complement each other. **H9**: Local is better than global for identifying false positive (FP) recommendations, and **H10**: global is better than local for identifying false negative (FN) recommendations. The hypotheses' associated metrics are provided in Table 2 and are described further in Section 4.2.

4.1.2 Participants and Treatments. In the same manner as in Study 2, thirty computer-science researchers were recruited and separated into treatments. A few changes were made to the explanations. Their titles were made purple and revised to better draw 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 described in Section 4.1.3, Study 3's procedure no longer required participants to curate recommendation feeds, so the only clickable buttons were for looking at explanations and abstracts. The remaining buttons were included to provide context for how the recommender would work overall. Furthermore, the LG condition was updated so that the local and global explanations were presented separately because their unified presentation in Study 2, in which local explanations could only be opened one-by-one, may have reduced focus on the local explanation. Study 2 participants in the L condition opened an individual local explanation 9.3 times on average, while those in the LG condition did so only 2.7 times on average.

Study 2 participants' feed topics varied largely in breadth and familiarity, which may have hindered our ability to observe significant results. Thus, Study 3 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 topics were chosen based on three criteria: 1) for researchers from various areas to engage with and understand the feed, the topic had to use limited jargon, 2) the topic had to be specific enough that FPs occurred, and 3) the topic had to be broad enough that a FN cluster emerged. For example, in the "misinformation on social media" feed, true positives (TPs) were exclusively about *covid-related* misinformation, so any papers discussing misinformation on social media unrelated to covid formed a FN cluster. Each feed was seeded with five carefully selected papers.

Three annotators classified the top 20 papers of each 250-paper feed as FPs or TPs and the bottom 50 papers of each feed as FNs or TNs, based on paper titles and abstracts. Only papers upon which there was unanimous agreement were added to the pool of papers that participants could encounter. The 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 TP papers for the study. Also, the "misinformation on social media" feed had ten FNs. Two did not belong to the FN cluster about misinformation on social media *unrelated to covid.* To ensure that all participants interacting with this feed would see a FN from the same cluster, the two FNs were removed from the pool of papers.

4.1.3 Procedure. Participants first opened a link to the recommender. 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 FP 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 FPs. About half of all the TPs 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 TP was represented accordingly, one such TP was randomly chosen to be included in each participant's feed. **H10** was addressed next with a FN survey.

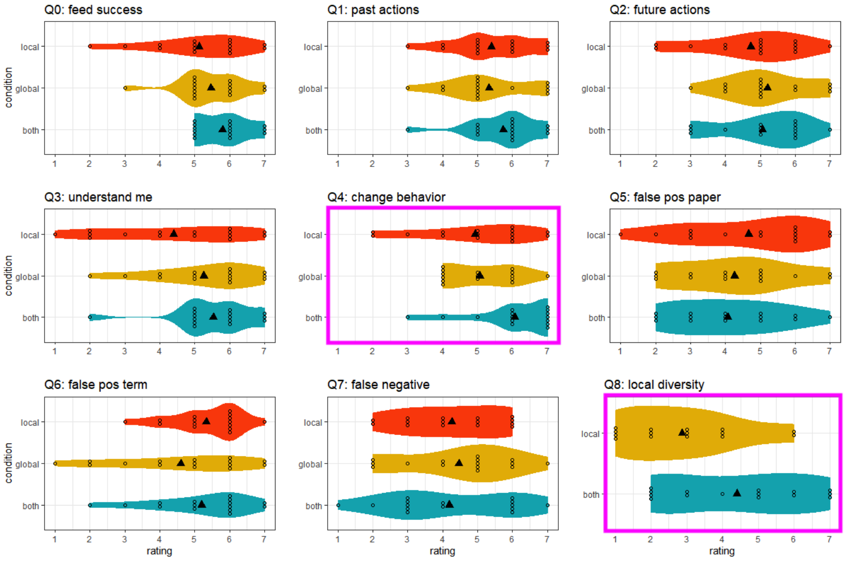


Figure 3: Study 2 results for each Likert-type question and condition. 1 means "strongly disagree," and 7 means "strongly agree." Triangles represent the mean responses for each question/condition. Circles represent individual responses. Q4: With both explanations rather than only global (p=0.015, uncorrected) or local (p=0.030, uncorrected), participants were more confident in explaining how they would like the system to improve. Q8: Participants desired less diverse local explanations when global was present (p=0.038, uncorrected).

Table 2: Study 3 metrics for hypotheses in Section 4.1.1. Scores are described in Section 4.2. Question is a 7-point Likert-type question.

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."

The survey presented participants with three new paper recommendations for the feed. Two were TPs and one was a FN. The survey asked participants to rank these papers based on how they believed the recommender system *would rather than should* rank them. Ideally, the participant would be able to recognize that the FN 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.

4.2 Results and Discussion

Here we discuss results for **RQ2** and **RQ3**'s hypotheses and metrics (Table 2). The FP survey score **M1** was calculated as follows. For each recommendation, 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 score. If they classified it incorrectly, -1 multiplied by their confidence was added. The FN survey score **M2** was 1 if the FN paper was ranked below the two TP papers and 0 if not. For **M1** and **M2**, we analyzed all condition pairs with a one-way ANOVA test. For **Q9**, we compared the L and G conditions with the within-subjects two-tailed Wilcoxon signed-rank test and the other condition pairs with the between-subjects two-tailed Mann-Whitney-Wilcoxon test. The significance threshold was p < 0.05. All pre-registered results were insignificant.

Regarding **RQ2**, the slight difference in wording between **H11**'s **Q9** and Study 2's **Q4** may have implied that, to respond affirmatively, the participant needed a technical rather than merely conceptual understanding of how the system could improve. Also, they may have had more trouble conceptualizing how it should improve, as they did not choose the feed topic. These points may explain the lower average response to **Q9** (4.67) as compared to **Q4** (5.36).

Regarding **RQ3**, while we did not find a significant difference among the conditions with respect to how well participants completed the FP (**H9**) or FN (**H10**) survey, we observed uncorrected significant differences among the conditions in terms of how *quickly* participants completed the FP (F(2,19)=5.216, p<.05) and FN (F(2,19)= 4.801, p<.05) surveys for the "misinformation on social media" feed, as illustrated in Figure 4. Twenty-two participants interacted with this feed (8 in the G condition, 7 in each of the other conditions). Time spent on each survey was rounded to the nearest half-minute.

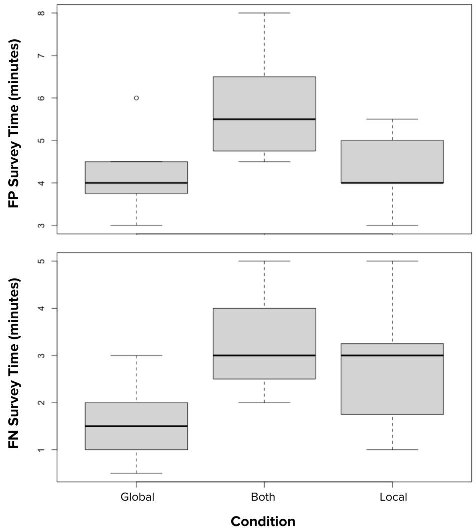


Figure 4: How much time Study 3 participants spent on the FP (top) and FN (bottom) surveys as a function of condition, for the "misinformation on social media" feed. Top: Participants spent more time on the FP survey when both explanations were present as compared to only global (p=0.020, uncorrected) or only local (p=0.045, uncorrected). Bottom: Participants spent less time on the FN survey when only global was present as compared to both (p=0.018, uncorrected) or only local (p=0.135, uncorrected). The global explanation alone thus appears more helpful than both explanations together for identifying FPs and FNs efficiently.

For these results, we analyzed all condition pairs with a one-way ANOVA test followed by a Tukey HSD test. These results are not further corrected because they were not pre-registered for analysis. Figure 4 shows that participants with the "misinformation on social media" feed in the LG condition completed the FP survey slower than those in the G (p=0.020,uncorrected) and L (p=0.045, uncorrected) conditions. This suggests that **providing both explanations rather than either alone causes users to identify FPs more slowly**, which may simply be due to the fact that there is more information to consider when both explanations are available. This result is not necessarily obvious, as having both explanations could have allowed participants to more easily recognize FPs. For the same feed's FN survey, Figure 4 shows that participants in the G condition completed it faster than participants in the LG condition (p=0.018,uncorrected), suggesting that **providing only a global explanation rather than both explanations helps users identify FNs more quickly**. Though insignificant, participants in the G condition also completed the survey faster than participants in the L condition (p=0.135,uncorrected). These results make sense for two reasons: 1) with LG, users have more information to evaluate, and 2) in comparison to L, G's top terms provide users a more straightforward indication of which terms the model may be considering too important or unimportant, which can cause FNs.

Reasons we did not see the same results for the "educational technologies for demographically diverse users" feed may include: 1) the topic appeared more difficult to understand, 2) the FNs resulted from an over-specification (for cultural diversity) as opposed to an unnecessary specification (for covid), and 3) the FN-related global terms were less prominent. However, in a follow-up formative study with time-constrained surveys, participants were not evidently better at identifying FPs or FNs in one condition versus another. Potential reasons are: 1) researchers are used to evaluating paper relevance without explanations, 2) identifying FPs and FNs may have been abnormally easy for the widely accessible feed topics, and 3) participants were not necessarily invested in the topics.

5 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 two explanation approaches' uses in this system. We saw 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 alone compared to when a global explanation is available. We also found qualitative evidence that, in the domain of research papers, local and global explanations may have different advantages with respect to research exploration. 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 positive and false negative recommendations. While we did not find differences between the two explanations in terms of user accuracy in identifying false positives or negatives, we did observe evidence suggesting that having both rather than either alone slows users' identification of false positives, and having a global explanation alone rather than both 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 (computer-science 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 if the recommendation itself summarizes its contents, as with papers. Future research may also explore how these explanations are used in higher-stakes recommendation settings such as education or healthcare, in which explanations often bear greater importance. Finally, future work may investigate how these explanations are used for purposes other than clarifying recommendation relevance, such as discovery of more diverse recommendations.

ACKNOWLEDGMENTS

This research was supported by the University of Washington WRF/Cable Professorship and the Allen Institute for Artificial Intelligence (AI2). The authors thank Matt Latzke and Cecile Nguyen for their advice regarding the project's user-interface design and code respectively. The authors thank all the participants and pilot participants who made this work possible. Finally, the authors thank the anonymous reviewers of this work for their valuable feedback.

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

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

[5] 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*.

[6] 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).

[7] 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.

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

[9] 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).

[10] 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).

[11] 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.

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

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

[14] 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).

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

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

[17] 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).

[18] 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).

[19] 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).

[20] 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).

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

[22] 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*.

[23] 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.

[24] 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).

[25] 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*.

[26] 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).

[27] 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).

[28] 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).

[29] 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.

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

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

[32] 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).

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

[34] 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).

[35] 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).

[36] 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.

[37] 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.