CREST: Mediating Collaborative Search and Agreement for Group Property Bookings - A Changelog

CREST was submitted to CHI 2023. In this document, we address how we have iterated on this work to tackle the reviewers' comments.

We note that CHI's reviewers found our system "well-implemented," "comprehensive," and "well-motivated." They found "the system presented [to be] novel and exciting" and they appreciated "the thorough/diligent design process that is clearly well-informed from the literature and backed up by the formative conversations/study." They found the work to be grounded "in concrete problems and approaches, which makes the final design feel very well-motivated." They commented that the "evaluation is realistic. I am convinced that the system was useful for the participants."

Major Revisions

1. A more detailed explanation of our system implementation details (Reviewer 1AC, 2AC) : We have included more details regarding the implementation of CREST. Specifically, we included a system architecture diagram that explains how our tool works end to end and have added a dedicated Implementation detailing the mechanics of our tool. Furthermore, we included a table in the appendix (Table 6) that explains how each CREST-Bot Messages is generated and what effects its corresponding action buttons do.

2. Additions to our design section (Reviewer 1AC, 3): We have expanded our motivating design section by including a table (Table 1) that provides an overview of how our identified tasks and challenges are tackled by the different features in our tool. We hope that this table further solidifies the motivating analysis and studies we did which motivated the creation of CREST.

3. Included a comparison's table between CREST and previous work to expand the literature review (Reviewers 1AC, 1 and 3): We included a comparison table (Table 5) that directly compares our tool with various tools surveyed and included the references that the reviewers shared with us. We still find that we are the first to place mediated agreement as a pivotal part of the search process.

4. Increased the participants' size of our user study to increase its validity, appended the study design, and included more details on the study's implementation, promise, and limitations (Reviewer 1AC and 3): We have included explanations on why our user study had to be designed the way it was and are upfront about the limitations the results represent. We also classified the messages in both tools in order to have a more granular understanding of how Agreement occurs in Baseline, a metric we failed to

capture before. We also increased the number of groups using Baseline and CREST and replaced some of the old ones from the first paper as they didn't contain sufficent logs to support some of the new analysis we do in this submission. Unfortunately, this led us to drop groups that didn't agree in the baseline. Nevertheless, we are happy to see that our general trends, results, and patterns stand in another run.

5. Included a future-work section to show how CREST can inform motivations in other field (Reviewer 1AC and 2AC)s: We are very excited about how the lessons learned from CREST can be used in future research directions, and included an entirely new section (Section 5) detailing how CREST can be used in new fields.

For each reviewer comment, we provide exactly the changes made to address it within the review text below.

******
Our replies are surrounded by ******.
******

Original CHI 2023 Reviews

1AC review (reviewer 4)

Expertise

Knowledgeable

Originality

Low originality

Significance

Medium significance

Rigor

Medium rigor

Recommendation

I recommend Reject

1AC: The Meta-Review

Thanks to the authors for the paper submission. While all reviewers saw potential in this work, the consensus was to reject this paper given the large amount of changes needed. Below I will synthesize feedback from the reviews but please look at the individual reviews. With improvements to the work, I encourage the authors to resubmit this work in a future venue.

KEY STRENGTHS
- The paper is relevant to the CHI community.
- The authors develop a working system that seems well-implemented and comprehensive.
- The authors undergo a thorough and diligent design process informed from prior literature and formative conversations. As a result, the design feels very well-motivated.

KEY WEAKNESSES
- The authors could better explain details of how their system works [2AC]. They should clarify what parts of the system are fully-functioning and what parts are manually supported. They should also clarify implementation details that would be useful for other researchers to better understand or build on the prototype System.

******
There were no manually supported components. We now describe in Section 2.4.1 the implementation and the preconditions that trigger CREST-Bot messages in Table 5 in the Appendix.
******

- The authors could cite more relevant work and explain in more detail what parts of their work are novel compared to prior literature in related domains [R1, R3]. It would also be useful to discuss why the choice to go in a similar or different direction compared to prior work.

******
Table 4 provides a clear summary of the relationships and differences between our tool and prior work.
******

Finally, given similarities in features with
prior work, the authors could consider doing a comparative evaluation with a

baseline that is closer to prior work so as to sharpen the evaluation to be about just the novel aspects of this system.

******

Our baseline is representative of current systems and it provides collaborative search functionality and chat features.

******

- Many details about the evaluation study are missing [R3]. In addition, a more in-depth analysis of the results could turn up additional insights about why certain outcomes were the way they were.

******

We added many details throughout the Evaluation section and a few more qualitative findings.

******

- The Discussion could be improved to talk about how the work can inform HCI researchers in the broader domain of collaborative search and consensus-making [2AC]. This ties in with the earlier point to discuss in more detail this work's difference from prior work in other domains - by being more in dialogue with literature in the broader domain, the authors can have a more informed discussion of their work's significance and relevance to broader design questions.

******

We added more details in the conclusion section to this point.

******

2AC review (reviewer 2)

Expertise

Knowledgeable

Originality

High originality

Significance

Medium significance

Rigor

Medium rigor

Recommendation

I can go with either Reject or Revise and Resubmit

Review

This paper presents a novel system called CREST that enables collaborative search, discussion, and agreement for rental property bookings. Overall I think the system presented is novel and exciting, and I appreciate the thorough/diligent design process that is clearly well-informed from the literature and backed up by the formative conversations/study. This really helps to ground the work in concrete problems and approaches, which makes the final design feel very well-motivated. However, I think there are a few important things missing for this to be a significant contribution to HCI.

These are: 1) a lack of information about how the system is actually implemented (which would make replicating it very difficult) and

******

See Section 2.4.1 for expanded details

******

2) a lack of discussion of how this system and the paper's findings could inform the broader domain of collaborative search and agreement beyond the specific domain of rental properties.

******

We believe that even targeting a specific domain like shared booking is a worthy cause. We motivate the importance of this application in the introduction. We also discuss other domains that can benefit from our design principles in the Future Work Section.

******

This might be okay if the evaluation provided a significant contribution on its own, but given its limited size and external validity, I'm not sure that there's enough here for other HCI researchers to confidently use and build on the results.

******

Qualitative studies of the form we conduct can inform future research and tool design. We show that providing agreement mechanisms through mediation empowers users engaged in collaborative booking tasks. Thus, our work is of benefit to future HCI researchers in that (1) it expands the design space by providing novel features such as contracts, house rules and mediation bot. (2) it provides a qualitative study that richly describes how users behave when presented with these features and what subtle effects can different mediation messages have on users (Section 3).

******

Revising the paper to address these changes would require significantly cutting down on the existing content (specifically, I think section 2 could be made much more concise), and I feel these changes would likely be too substantial to make within the 5-week R&R timeline. However, I do think the paper has a lot of strengths (well-motivated design process, thorough reviewing of prior work, clearly-presented results) and would love to see it published in the future.

1) System Implementation:
The user experience of the system as well as the reasoning for its design decisions are very well-described in section 2. However, I was left with many questions about how the system actually works.

******

See Section 2.4.1

******

- How does the system decide when to show a notification, and what notifications to show? Pg. 9 states, "Simple rules control the generation (and disappearance) of CREST-bot's messages." If these rules are simple, it would be helpful to include them in the paper.

******

See Table 5

******

- Where does the system get its underlying property listings from? E.g., an online database, or an existing property listing website? If it was manually populated for the study that is fine, but some discussion of what it would take for this to be a usable system in the real world would be helpful, as that is likely a reason why people currently use multiple disconnected tools - because each tool has data that the others do not.

******

See Paragraph 1 in Section Section 2.4.1

******

- What are all the different possible actions and how do they work? E.g., what does "reallocate contributions" do? (Table 2 #7)

******

See Table 5 in the Appendix

******

- How are the collabo-ratio factors calculated? Specifically, I'm a bit confused

about the “influence” factor. The word influence seems to imply that a user with a high influence score would have more influence over the final decision / contract than other users, but the definition (“the user’s interaction with other users’ preferences or contracts”) seems a bit different. Some more clarity around the meaning and operationalization of these terms would be helpful.

******

See Paragraph 6 titled Collaboratio in the Awareness Principle of Section 2.4. These metrics are saved as simple counter variables for each user in a table in our database and are increased anytime the user meets any of the corresponding conditions are met. *Activity* or the contribution of the user to search preferences and contract proposals, *Influence* or the user's interaction with other users' preferences or contracts, and *communication* or how often the user chats with others.

******

I understand that the intended main contribution of this paper might be the design process itself, rather than a fully-functioning system. However if so, I would expect the paper to be upfront about what parts of the system are fully-functioning, and what parts are manually supported. Currently it sounds like CREST-bot is fully automated, but a description of how the automation actually works is missing.

******

All are fully automated, there were no manual components in the prior submission and none now.

******

2) Future work / applicability to other use cases
This paper targets a fairly specific use case, and does it quite well. But in order to make a significant contribution to HCI, it should at least discuss how the proposed approach might (or might not) apply to other scenarios beyond rental booking. Are there other use cases involving collaborative search & agreement that could potentially benefit from these design ideas? This is addressed very briefly in the last sentence of the conclusion, but given how thoroughly this paper discussed prior work in the broader domains of collaborative search and conflict mediation, it feels like it is missing a broader discussion at the end that could tie the specific findings from the evaluation back to these domains, to help inform future work in the space.

******

We included an entire section on how CREST can be applied in the fields of finance and investing, event planning, and charity.

******

Other notes:
I thought that the evaluation section was well-structured, and the results were

clearly explained. I just have one concern about how the task is described in section 3.1. While section 3.2 is upfront about the study's limitations, section 3.1 seems to over-claim the external validity of the task. Specifically, I think it's a bit of a stretch to call the task of being a personal shopper for a fictional avatar "a commitment akin to that of an enthusiastic tenant-to-be rigorously searching as they are to make a substantial time and financial investment" (pg. 11). Of course this is a difficult situation to replicate for an experiment, so I think it's fine to use a fictional situation instead, but the paper should acknowledge this limitation rather than claim it is comparable to a real-life scenario. Especially since the participants were anonymous to each other, this is a pretty different scenario than what the tool is designed for. A group of friends or family members would have existing personal relationships with each other that could affect how they interact in a tool like this.

******

See Pararagraph 4 in the evaluation section (Section 3.1) that justifies our approach

******

---

reviewer review (reviewer 1)

Expertise

Knowledgeable

Originality

Low originality

Significance

Medium significance

Rigor

High rigor

Recommendation

I can go with either Reject or Revise and Resubmit

Review

This paper describes a system called Crest, which assists a small group of users in negotiating and deciding on different lodging or co-living options. The authors conducted formative studies and literature reviews that motivated the system's design well. The system from the video looks comprehensive and well-implemented, and the evaluation is realistic. I am convinced that the system was useful for the participants.

The core issue I had while reading this paper is that the claim of "No other tools that supported all three stages of search, discuss and agree" is a little too strong. More specifically, the following two papers deal with a similar context of small groups picking restaurants:

Hong, Sungsoo Ray, Minhyang Suh, Nathalie Henry Riche, Jooyoung Lee, Juho Kim and Mark Zachry. "Collaborative Dynamic Queries: Supporting Distributed Small Group Decision-making." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

Hong, Sungsoo, Minhyang Suh, Tae Soo Kim, Irina Smoke, Sangwha Sien, Janet Ng, Mark Zachry, and Juho Kim. "Design for collaborative information-seeking: Understanding user challenges and deploying collaborative dynamic queries." Proceedings of the ACM on Human-Computer Interaction 3, no. CSCW (2019): 1-24.

******

We refer to these papers in our related works and explain our relationship and differences with them.

******

For example, they also assumed users have multiple semi-structured preferences when picking restaurants together (such as price, ratings, and item preferences). The system proposed in prior work included designs where users could see each other's preferences (including ordinal, range, and binary preference types) and use a chat window to negotiate the final decisions. These are not currently discussed and contrasted with the current work, making it difficult to determine which designs were novel and different.

******

Our work is motivated by previous work and this particular paper mentioned only motivates our Collaborative Query Panel feature. This was done to match the state of the art in collaboration and search and our tool innovates in mediating agreement effectively through our contract and signing components and our automated mediation agent CREST Bot. See Table 4 for a more in-depth comparison of our tool and previous work.

******

Quickly re-reading the above papers, it seems like chatbot mediation is one thing that was significantly different from prior work (besides different but similar

contexts of restaurant vs. lodging). The contrast authors made between algorithmic arbitration and social chatbot mediation is interesting. The authors also cited prior work, which can motivate the design choice. In prior work above, they allowed users to manually "ping" others for agreement instead. A careful discussion of the differences will be important for the potential revision. In the discussion, it would be nice to see the authors' stance on whether incorporating algorithmic arbitration into a system like Crest has potential benefits or intuitions about why this might not be a good idea.

Given how similar some of the designs seem to the system in prior work, I wonder if it makes more sense to only disable the features unique to Crest as a more substantial baseline. Such as the chatbot mediation feature and the Collabo-ratio Visualization.

******

But also contract and house rules are unique to CREST and the baseline contains chat and collaborative query search and shared bookmarks which are typical of collaborative search tools.

******

---

reviewer review (reviewer 3)

Expertise

Knowledgeable

Originality

Medium originality

Significance

Medium significance

Rigor

Low rigor

Recommendation

I can go with either Reject or Revise and Resubmit

Review

This paper explores a system solution for a collective information search and group decision-making in group property bookings. The novel contribution is in the goal-centered flow design principles that shepherds users to complete the goals with a contracts list and some nudges from a mediating chatbot.

I found the angle of driving the group to make a decision as soon as possible interesting, but I doubt whether that is always an ideal outcome for the group. A pre-mature decision closes the door to exploring a better outcome. It would be great to see the discussion on balancing this trade-off in the paper.

******

We do not see evidence of CREST's goal centered flow to be minimizing individual search or exploration or driving the group to make a decision as soon as possible. In fact we note that teams create multiple contracts and discuss them collectively and spend the same amount of time collectively on both CREST and the baseline. What CREST does is center the user's attention on finding a workable property instead of purpose-less open-ended search. See the last Paragraph of Section 3.2.2.

******

The paper could add to the knowledge of the HCI community. The application of the theoretical mediation framework on the bots and the report on how people responded to the mediating bots could be cited by other future papers. (Disclaimer: I only have passing knowledge of Chatbot interaction). The proposed system has various novel components (e.g., liking/disliking a preference).

Yet, the paper seems to lack depth.

The author(s) conducted a small needs-finding study to understand the tasks and the challenges better. However, there are not many details on the study itself. It is unclear who the participants were and how they were recruited.

******

See Paragraph 4 in Section 3.1 for details. 21 people were recruited for CREST and 21 people were recruited for baseline from the researcher's university campus. We provide a demographic breakdown of the participants in this paragraph too.

******

There are also

no details on how the study was conducted, what type of data were collected, and how the data were analyzed. The lack of such details raises a question of the

validity of the needs-finding study.The analysis of tasks and challenges does not make concrete references to the needs-finding study. While synthesizing the observations and existing literature into principles is important, it is more convincing to see how the design decisions are also driven by real observations. The author(s) could consider providing supporting examples/quotes from the needs-finding study to the Task Analysis and the Challenges sections.

******

We had an open-ended interview about their experiences with a collaborative booking task.. Our tasks and challenges are also motivated by the literature review and initial pilot runs with different prototypes. The design section distills all of these real findings.

******

The results of the main study are quite rough. A more in-depth analysis of the results could reveal some extra insights. For example, the paper only reports agreements in the main study but doesn't specify the quality of the agreement beyond subjective ratings from the participants. I would be curious to see whether there was any case where some avatar requirements weren't satisfied. There is no analysis of the chat messages between participants to see the kind of communication that happened. The chat messages analysis could reveal more insights into why some of the Baseline groups didn't reach an agreement.

******

See Paragraph 4 in Section 3.2.1. In CREST, the average number of satisfied requirements was 2.52 (σ = 0.46) and in Baseline it was 2 (σ = 3.06). We also find more agreement activities in CREST than in baseline even after doing a natural language classification of the messages as seen in Figure 6.

******

There are also missing related works on collaborative dynamic queries ["Collaborative Dynamic Queries: Supporting Distributed Small Group Decision-making" (CHI 2018), "Design for Collaborative Information-Seeking: Understanding User Challenges and Deploying Collaborative Dynamic Queries" (CSCW 2019)] that shows how a group can search for alternatives with awareness of others' preferences while making a decision together. The author also didn't mention some related "search integration with other applications" like "SearchMessenger: Exploring the Use of Search and Card Sharing in a Messaging Application" (CSCW 2017).

******

Done. This papers have been included in Section 4. See Table 5 for more details.

******

Presentation-wise, the writing is clear and easy to understand. However, the organization of the paper could be improved. It is hard to keep track of the challenge(s) each design principle addresses. The author(s) could show the

addressed challenges in italic/bold fonts to clarify the connections. Another approach I have seen is numbering the challenge (e.g., C1, C2 ..) and referring to those in the following text.

******

Done. See table 1 and how we now specifically provide a table that relates each design principle and feature to a task or challenge.

******

Further, the broad actions categorization (search, discuss, and agree) is muddy. The boundary between "discuss" and "agree" blurs regarding negotiation. Without a clear distinction between actions, the follow-up results and analysis based on this categorization become less reliable.

Although making this paper stronger wouldn't require a re-run of the study, it would involve an extra analysis, a significant reframing, and a rewrite.

Minor notes:
- The Related Works section appears late in the paper. Some other communities put the Related Works section near the end. This is not a dealbreaker. Just a note that I typically see the Related Works section near the beginning of CHI papers.
- 532: I am not sure how shopping for someone else would make the participants more committed.
- The statements at the beginning of the Results section could be put into their own separate "Limitations" section.
