Online Crowdsourced Questionnaire

1) Give the final version of your survey both as it appears to the participant and in a format that we can succinctly read.
Screenshots of the final version of the survey are provided as separate file 'Survey Final Version.pdf'
Survey link: https://utah.sjc1.qualtrics.com/jfe/form/SV_0qBWD8RrM1XaXmS

2) For each questionnaire question, give a detailed explanation for why you are asking it.

Introduction:

1. What platforms do you use that provide you with personalized content? (list of platforms, multi select)
    a. Gives you an idea of what kinds of apps they use. That way, we can ask questions specifically based on the platforms/categories they select.
    b. This will give us an idea of how to compare between platforms, but also categories (social media vs entertainment streaming).
    c. By finding out with respect to platforms, we will be able to assess trust and safety with respect to different platforms.
2. Imagine that data about preferences (what movies you watched/songs you listened to, your social media history (likes, follows, interests), ordering history) from all the platforms you use was leaked, would all data harm your personal safety equally? (Yes or no)
    a. Just trying to find out whether all data is the same, and whether the user would prioritize some type of data over the other).
    b. It's relevant because the type of data directly ties in with the perception of safety and trust.
    c. It's best left as a yes or no question because all we're trying to find out is whether all the data is the same (without any nuances.)

Specific for each platform that they clicked on question 1 in the Introduction (at most three):

1. I trust *platform* with handling my personal data and information with respect to my privacy.
    a. We are trying to find out how the participant perceives their privacy is respected by the recommender system.
    b. It directly applies to our research question.
    c. The question includes an introduction of the scenario and we ask a question about the level of trust. Answered on a five-point scale.

2. How much would your personal safety be at risk if any of the historic or personal data on *platform* would be subject to a data leak?
    a. We are trying to determine if the participant is aware of any data that might be compromising their personal safety.
    b. This question is related to trust in the recommender system.
    c. We are asking for a level of risk which can be answered on a scale.
3. When using personalized recommendations from *platform* do you feel the recommendations are biased in any way?
    a. Trying to find out if the users feel a bias in the recommendations provided by the system.
    b. This question tries to establish if the users' trust in the system is affected by the biased content it provides.
    c. The scale allows us to find out the level to which the user feels bias in the system.
4. How satisfied are you with the recommendations system provided by *platform*?
    a. This question finds out the satisfaction level of a user in the system.
    b. This question tries to find out the engagement (potential trust x usage) level of the user with the system based on satisfaction of content provided.
    c. The scale directly addresses the satisfaction a user is having with recommendations in general.
5. How much influence do you think the recommendations from *platform* have on your time spent on the platform?
    a. This question finds out the usage perception of a user because of the recommendation system.
    b. This question ties in to the comfort level of the user with the recommendation system.
    c. It quantifies the users perception of usage caused by the recommendations.
6. How much influence do you think the recommendations from *platform* have on your money spent on the platform?
    a. This question finds out the spending perception of a user because of the recommendation system.
    b. This question ties in to the comfort level of the user with the recommendation system.
    c. It quantifies the users perception of spending caused by the recommendations.

Demographic questions

1. Please select your age-group
2. Please select your gender
3. Please select your marital status
4. Please select your occupation
    a. We are trying to find out the general demographics of our participants.

b. When doing the analysis, we will see whether any of the demographics are a major factor in their perception.
c. Most of these would just be one answer, which is how demographics generally are. We have excluded race. We are also using the answers as additional quality measures by cross-checking the answers from the survey with the demographic information provided by Prolific.

3) RQ: How do personalized recommender systems affect human behavior with respect to trust and safety according to user perception across different platforms?

4) Advantage of online questionnaire method
Online questionnaire method is a great way to quantify users preference according to their perception. It also eliminated any bias or discomfort that might be felt in interviews. The crowdsourcing platform gave us quick access to a diverse user base. We can not do this via interview or log analysis because even though we can quantify user preference using log analysis we cannot get user perception from it. Similarly for interviews we might be able to get user perception but not quantifiable insights plus. Online questionnaires are able to give us unbiased, quantifiable and structured insights. Especially since we are trying to understand user perception towards personalized recommendations and get quantifiable results, online questionnaires are a perfect fit.

5) Yes, the task results did help us answer the research question. We received enough data to quantify user perception towards personalized recommendations. We also were able to find out some applications which the participants use frequently. We changed the question about whether the platform affects the participants' resource consumption (time, money) into two separate questions in order to get the granularity for the additional insight.

6) Insights gained from spending time as crowdsource worker
Time spent in Prolific as a worker included the qualification tasks and participating in a few surveys. This was useful as to see how the attention questions and required information like the unique Prolific ID and the survey code are handled. On Prolific, the participants provide their unique Prolific ID so that the survey responses can be correlated with the Prolific results. Prolific also provides some basic demographics like gender, age, employment status regardless. This can be used as an additional measure to validate the survey responses. The attention question in our survey uses the main idea from another survey taken on Qualtrics. Also, providing the survey code in the completion message which is displayed after all survey questions have been

answered ensures that participants complete the survey and don't abandon the survey inadvertently as they think that they may have already completed the survey.
One group member also spent time on Amazon Mechanical Turk, but found that it was kind of difficult to find legitimate tasks. Most of the tasks would lead to external surveys that would lead to another website with yet more surveys (through which you get paid, it was strange). Because there was so much spam, it was difficult for an untrained eye to spot spam from others. There was no review of the entity assigning tasks to be found, so you would have to attempt to complete the task before finding out whether it is legitimate or not.

7) With how many workers did you pilot your task? Why?
We ran a pilot with five participants. We chose a smaller number of participants so that we would still have sufficient participants for the full survey. We also wanted to see if the time we estimated was sufficient and the reward was attractive to receive responses quickly. For the five minute task, we offered a $0.67 reward, which is the minimum reward ($8.04 per hour). The tasks were completed within 30 minutes after publishing. The average completion time was 4 minutes 21 seconds. After the trial run, we split one of the platform related questions into two in order to get more granularity, essentially adding two to three questions per participant. Therefore, we continued with the 5 minute average completion time and the same reward for the full survey run.

8) What, if anything, changed in between your pilot(s) and your actual task deployment? Why?
We changed the question about whether the platform affects the participants' resource consumption (time, money) into two separate questions in order to get the granularity for the additional insight.

9) In what ways did you attempt to confirm that workers produced high-quality answers to your crowdsourced task?
We used three methods of confirming that the workers produced high-quality answers.
First, we added an attention question that the participants had to answer and select predefined answers and enter a word in a text box.
Second, Prolific provides some basic demographic information for each of the participants by their unique Prolific ID, which we use to correlate with the answers to the demographic questions we had at the end of the Qualtrics survey.
Third, we used the Qualtrics Q_RecaptchaScore for bot detection. The minimum score of our participants was 0.9 (>= 0.5 means that the respondent is likely to be a human).

10) What title, description, and keywords did you give your crowdsourced task? Why?
We kept the title simple: `Recommender System Trust and Safety` but we provided more details in the description: `We are conducting an academic survey about Recommender System Trust and Safety. We need to understand your opinion about safety related to personal information and how much trust you put into different platforms to keep your personal information safe.`

11) What was your reward for a successfully completed task? Why?
The reward for $0.67 for the 5 minute task, which resulted in earnings of $8.04 per hour. $8.00 per hour was the minimum we were allowed to pay. We wanted to maximize the number of responses we could get within the $25 budget. With this, we were able to get five responses for the trial run and 22 responses for the full run.

12) What other settings (e.g., number of assignments per task, time allotted per assignment, task expiration, task auto-approval, task visibility) did you use for your task? Why?
We selected participants from the US only because we would not have enough participants to cover a wider geographic area and get representative responses for each country. We allowed 30 minutes per task (which is the maximum on Prolific) in order to not rush the participants. We selected the standard sample of available participants in order to get the responses as quickly as possible.

13) Did you require any qualification criteria? Why?
Another requirement was to be fluent in English. We wanted to make sure the participants fully understood the questions.

14) Did you manually approve any tasks, or let Prolific auto-approve? Why?
We manually approved all the tasks in order to be able to validate the responses.

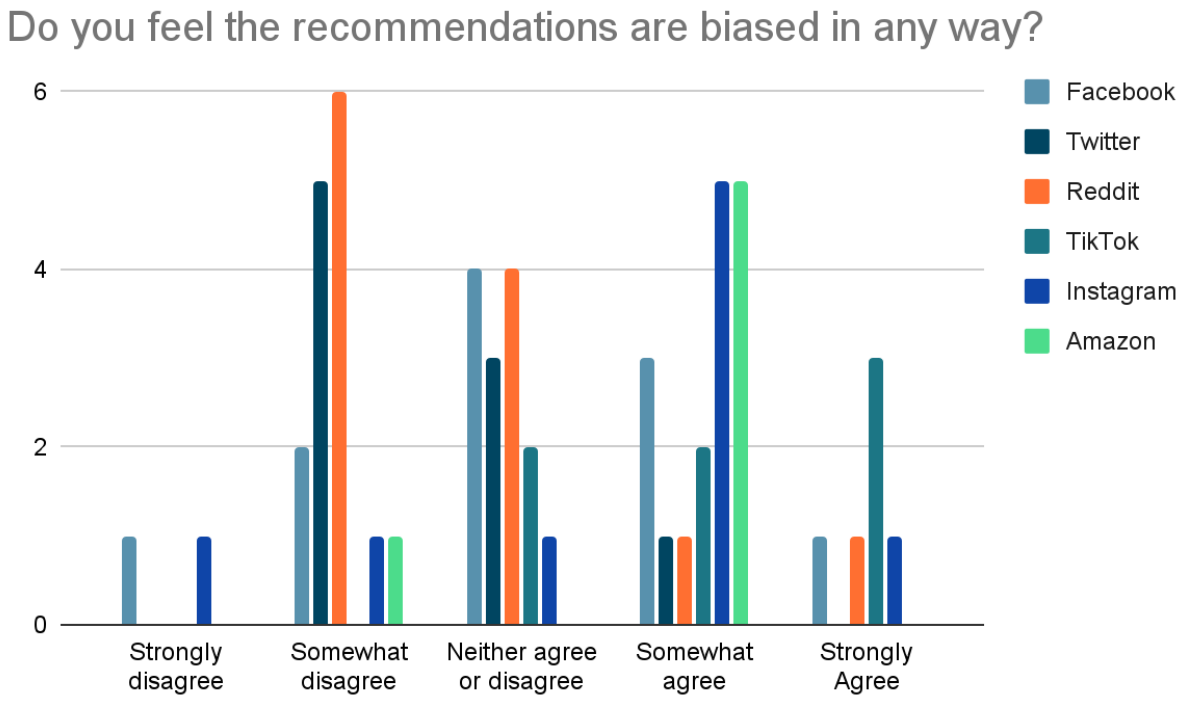
15) Did you reject any task submissions? Why?
We did not reject any of the responses. Although two participants only completed the attention check partially, we think that the wording in the question may have been ambiguous. They each provided two of the three of the required items correctly though. Also, the demographic information from Prolific matched their answers to our demographic questions in the Qualtrics survey.

16) Write a subsection for your final report on your online questionnaire method.
After the interviews were conducted and the data analyzed, a questionnaire was drafted based on the responses from the interview data. The main purpose of this questionnaire was to evaluate whether the findings could be generalized to a broader population. A pilot survey was conducted which got five responses on Prolific. The survey results that were analyzed got 22 responses on the same platform, with participants earning $0.67 for a single response which took about five minutes. General context was given as part of the attention check question, which the authors deemed to be cleared by all the participants.

17) Draft a subsection containing the questionnaire results for your final report (500 words).

After the analysis of the interview, the authors selected certain trends that were visible from the interview data. Based on that, a survey was designed and conducted on Prolific as a way to see if the trends observed during the interview data were generalizable to a more general audience. In particular, the authors were looking for trends regarding general bias, data privacy, and safety with regards to time or money spent because of the personalized recommendations. In order to get more granular information, participants were asked the same 6 questions for every app they selected.

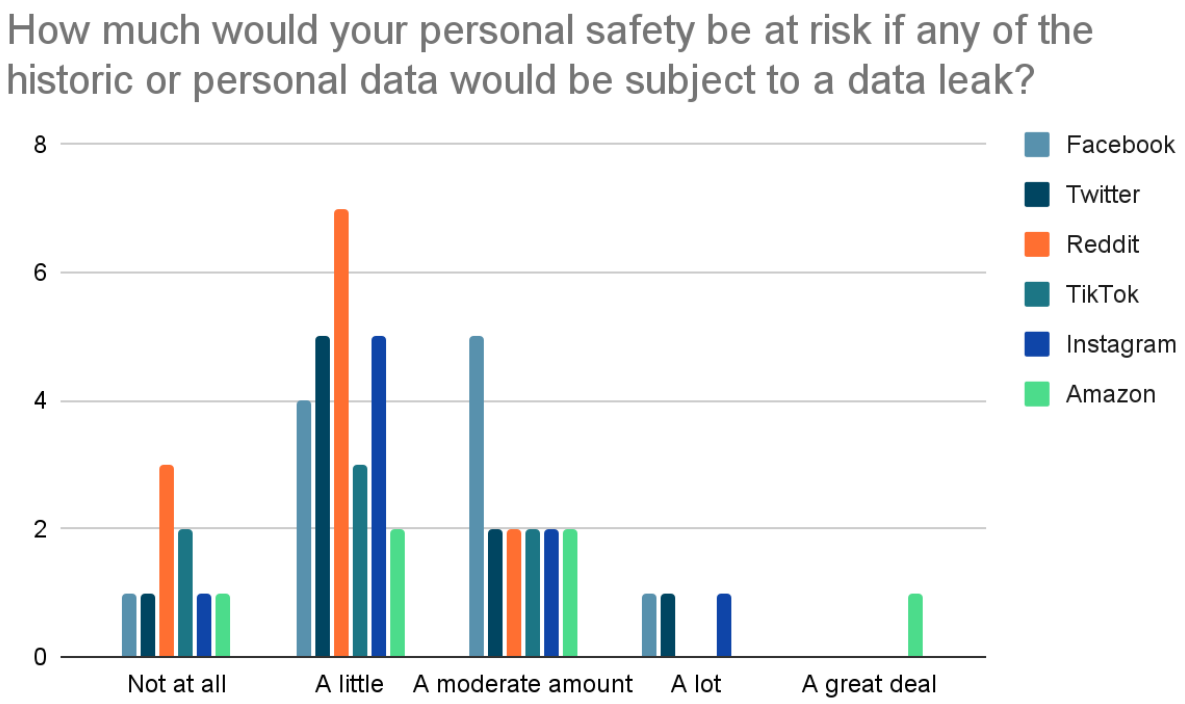
Bias:
Breaking down the analysis on a single app basis, the survey participants felt that some apps were much more biased than others. According to the participants, apps such as Instagram, Amazon, Facebook, and Tiktok were considerably more biased than others. This finding was largely consistent with what the interview participants had said with regards to the topic. The only part that cannot be corroborated in this case is what each survey participant perceived the definition of bias to be. This is particularly interesting because, purely by observation, these apps very aggressively utilize personalized recommendations as part of their in-app experience as compared to others. Twitter has both a home (recommended) feed and chronological feed as options, but it must be pointed out that even home feed only orders tweets by accounts followed (and a few promoted tweets in between). Reddit also utilizes membership in subreddits to curate content for the user. Based on that, it can be inferred that because of the perception of bias, users don’t trust a personalized recommendation heavy curation system. This was only based on 22 participants (with even less participants for single platform analysis), and a survey with more participants needs to be conducted in order to reach that conclusion.



Data Privacy

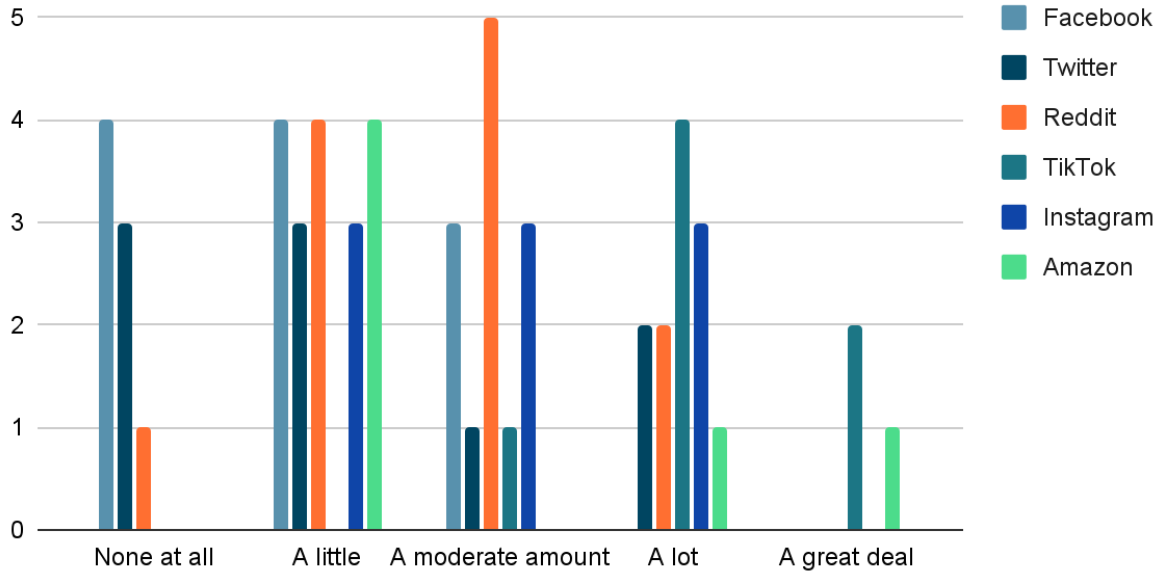
During the interview, a lot of participants stated that, in most cases, their personal safety would not be at risk if there was a data leak that happened. The exceptions were largely centered around large conglomerates, such as Amazon and Meta products (Facebook and Instagram). This finding was partially corroborated by the survey that we took. For Amazon, 50% of respondents stated that there would be a moderate risk or greater if there was a data leak, and 6 out of 11 said the same for Facebook. However, only 3 out of 9 would say the same for Instagram (which is the same as Twitter). That is particularly interesting because Facebook and Instagram are owned by the same company, so this finding can also suggest that context knowledge has a lot to do with

perceptions.

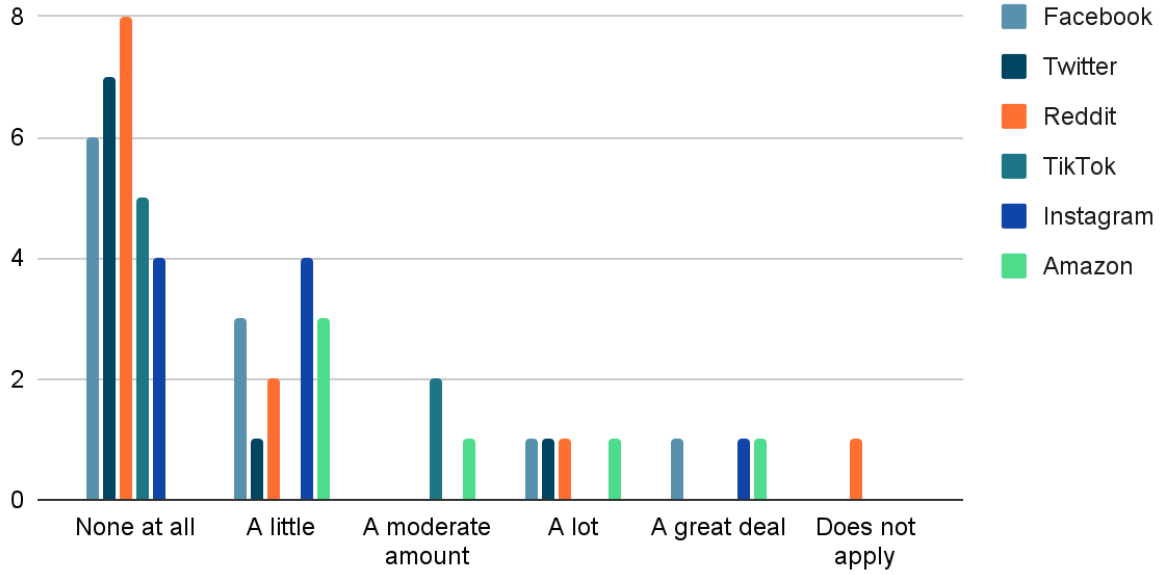


Time and Money consumption:
Majority of Tiktok and Instagram users pointed out that those apps increased their time spent on that platform as compared to others listed on our survey. This was a relatively new finding as this did not come up very often in interviews. Once again, it is an interesting finding that most participants said this about two platforms that very heavily rely on personalized recommendation for their in-app experience. This goes in line with Seaver's captive theory that the primary reason for personalized recommendations is to keep the users hooked in the app. A study with more participants is necessary to further substantiate this claim.

How much influence do you think the recommendations have on your time spent on the platform?



How much influence do you think the recommendations have on your money spent on the platfrom?



18) Did you use other resources? If so, what were they, and were they helpful?
We used some of the online documentation provided by Qualtrics and Prolific. They provided useful information about some of the advanced features, like the bot detection on Qualtrics and the tool that computes the exact cost for the study including fees on Prolific.