Pre-analysis plan: Recommender Systems and Consumer Choice - Experimental Evidence

Felix Schleef*

July 12, 2024

1 — Introduction

This project studies recommender algorithms as they are employed in e-commerce settings. Both in the US and in Europe, lawmakers are concerned with the power of large online platforms to influence users through recommendations. Current legislative approaches, for example the European Digital Markets Act, prohibit certain patterns of abusive behavior, such as self-preferencing (2022). I study an alternative legislation that would enable consumers to choose between recommender algorithm themselves. The idea that recommender systems could be chosen by consumers independently of the sales platforms was considered very early on in the economic literature on recommender systems (Resnick and Varian, 1997).

Existing studies on the effects of personalized recommender systems (Ursu (2018); Donelly et al. (2023)) use browsing data, where one set of users received recommendations by a recommender algorithm and another set did not receive recommendations. While it is appealing that these studies use data from real consumer choices, there are two key differences to the setting that I study experimentally: 1)

*CREST, GENES, Institut Polytechnique de Paris, Télécom Paris, Email: felix.schleef@gmail.com

The change in the platform design is not communicated to the users and they can therefore not change their search behavior accordingly. 2) Empirically, these studies are limited in the algorithms that they observe - they only observe the platform algorithm and a non-personalized algorithm.

In this project, I use an online experiment to study recommender systems. In a first step, I document the choice patterns of experimental subjects when their choices are ordered by different algorithms. Importantly, I address some of the gaps in the literature by informing subjects about the algorithm that creates their recommendations, and I introduce a consumer-optimal algorithm in addition to a "steering" (or "platform") algorithm.

Secondly, I study how subjects are choosing between different algorithms and elicit their willingness to pay for them. These results are aimed to inform a potential policy where consumers are given the power to change the recommender system.

2 — Conceptual Framework

2.1. Consumer choice

I follow Reimers and Waldfogel (2023) in modeling a subject's utility of choosing a lottery j ranked at r_j as:

$$u_{ij} = \delta_{ij} + \gamma_i r_j + \xi_j + \epsilon_{ij}$$

where γ_i denotes the user-specific effect of rank on utility, ξ_j is are unobserved lottery attributes (for example, degenerate lotteries might be more salient for subjects, leading to them being chosen more frequently), and ϵ_{ij} is an extreme-value error. Lastly, δ_{ij} denotes the rank-independent component of utility, which I parameterize as a CRRA-utility function:

$$\delta_{ij} = \sum_k p(c_k) \frac{c_k^{1-\omega_i}}{1-\omega_i} \tag{1}$$

where c_k are outcomes of lottery j, and ω_i is the individual specific risk aversion. Given that there is no outside option in the experimental setting, the choice probability for a lottery j ranked at r_j can be computed as:

$$P_{ij} = \frac{e^{v_{ij}}}{\sum e^{vij}}$$

with $v_{ij} = \delta_{ij} + \gamma_i r_j + \xi_j$. The ranking of a lottery affects both the subjects utility, and the probability that the lottery will be chosen.

2.2. Platform ranking

During the experiment, subjects make choices from lists of lotteries where the lists are created by two algorithms: the Expected Utility and the platform algorithm.

- *Expected Utility:* The expected utility algorithm orders the set of lotteries in descending order of CRRA-expected utility as in equation 1 and the top 10 lotteries are shown to the subject in order.

- *Steering:* The platform algorithm orders the set of lotteries according to the following index and the top 10 lotteries are shown to the subject in order of the index:

  $$I_j = \kappa_1 S\left(\delta_{ij}\right) - \kappa_2 S\left(\sum_k p(c_k) c_k\right)$$

  where $S(.)$ is a min-max scaler, and κ_1, κ_2 are the weights that the algorithm places on the objective of the subject and the "platform" respectively. This specification of the algorithm is inspired by Reimers and Waldfogel (2023). They use the index $I'_j = \kappa_1 ln(p_j) + \kappa_2 \delta_j$ and derive the Pareto-frontier between the consumer-optimal ($\kappa_1 > 0, \kappa_2 = 0$) and the seller optimal index ($\kappa_1 = \kappa_2$). For the purpose of the experiment, I will set $\kappa_1 = \kappa_2 = 1$ in the spirit of their seller-optimal index.

Throughout the experiment, the recommender algorithms serve two roles:

- Selection: Not all choices will be available to the subjects in each round. In each round, there is a set of 20 available lotteries. Of these 20 lotteries, only the 10 lotteries that rank the highest according to the algorithms criterion will be displayed to the subject.

- Ordering: The steering and the expected utility algorithm order the lotteries by their corresponding criterion.

3 — Experimental Design

The experiment features three kinds of tasks. The first task is a risk elicitation task in the style of Johnson et al. (2021), where subjects have the prospect of gaining either the outcome of a lottery that pays 10$ with probability 50% and 0$ with probability 50% or a sure but unknown monetary amount X, that has been randomly drawn with equal probability between 0$ and 10$. Subjects are asked to declare a threshold such that they would prefer receiving X if it is above the threshold, and they would prefer receiving the lottery if X is below the threshold.

The second task is a choice task, where subjects are asked to choose from lists of three-outcome lotteries that have been ordered and selected by the platform or the expected utility algorithm. These algorithms take into account the subjects risk aversion that has been calculated using the subjects choice in the first task and assuming constant relative risk aversion[1]. The outcomes of all lotteries are 10$, 5$, and 0$, and the respective probabilities are drawn to ensure that the probability of 10$ and 0$ does not exceed 50%. Throughout the choice task consumers are aware which algorithm has created the lists of lotteries, but the algorithms are displayed to them as "Algorithm 1" and "Algorithm 2".

Finally, I elicit subjects willingness to pay for one algorithm over the other. For

1. Since the two parts of the platform algorithm exactly cancel each other out in case the subject is risk neutral, subjects that are calculated to be risk neutral are randomly treated as either very slightly risk averse ($\omega = 0.01$) or slightly risk seeking ($\omega = -0.01$).

this step, I give subjects an endowment of 1$. Using a slider, subjects can change the probability that they will face either algorithm in the next set of 10 choice rounds. Choosing equal probability between the two algorithms (50% - 5ß%) is costless. Changing the probabilities by 1%, i.e. to 49% - 51% costs 0.02$. The cost increases linearly, so that the price of setting the probability to have either algorithm for the last 10 choice rounds is equal to the full endowment - 1$.

Timing:

- t=1
  - General Instructions about the experiment
  - Instructions on the risk elicitation task
  - Instructions on the lottery choice task
  - Instructions on the Willingness-to-pay (WTP) elicitation mechanism[2]
- t=2
  - Risk preference elicitation task
- t=3
  - 10 rounds of choice task with platform or expected utility algorithm (order of algorithms is randomized)
- t=4
  - 10 rounds of choice task with platform or expected utility algorithm (order of algorithms is randomized)

2. As suggested by Plott and Zeiler (2005), subjects will be exposed to training and practice rounds with the elicitation mechanism in an anonymous fashion. The practice rounds will however not be incentivized, as this is at odds with the recommendation of Azrieli et al. (2018) of selecting one round to reward at random.

- t=5
  - Willingness-to-pay elicitation
- t=6
  - 10 rounds of choice task with either random or personalized algorithm, depending on random draw with probabilities determined in the WTP-elicitation round

3.1. Incentives:

I divide the experiment into three parts:

- Risk elicitation
- First 20 choice rounds (10 rounds with platform algorithm, 10 rounds with expected utility algorithm)
- Willigness-to-pay elicitation and final set of 10 choice rounds (with randomly selected algorithm based on probabilities set in the willingness-to-pay elicitation step)

One of the parts will be randomly selected. In case the randomly selected part is part 1, subjects get a bonus according to their choice. In case the randomly drawn number X was above their threshold, they receive the X as the bonus, otherwise they receive the outcome of the lottery that pays 10$ with probability 50% (and 0$ with probability 50%).

In case the randomly selected part is part 2, I select 1 of the 20 rounds at random and the subject gets the outcome of the lottery they chose in that round. For example, a subject may have chosen the lottery [P(10$)=0.03, P(5$)=0.9, P(0$)=0.07] in the randomly selected round. The outcome drawn based on these probabilities is the subjects bonus payoff.

In case the randomly selected part is part 3, I add the unspent endowment from the willingness-to-pay elicitation to their bonus payoff. Furthermore, I draw one of the last 10 choice rounds at random and the outcome drawn based on the probabilities of the lottery that they chose in that round is added to the bonus payoff as well.

Comment on incentivization: The literature on rewards in experiments points out, that when multiple lotteries are paid out, the subjects might realize that risk is diversified and exhibit higher risk appetite than in other situations (see Azrieli et al. (2018) for a survey). I minimize this problem by randomly selecting one part to reward.

4 — Hypotheses

The primitive dataset consists of:

- Each participant's choice in the risk-preference elicitation task.
- Each participant's choices in the 30 choice tasks, the rank of their choice, the algorithm that created the list of lotteries, as well as the full ordering of lotteries that was presented to subjects in each round.
- Each participant's elicited willingness-to-pay

Using the primitive dataset, I further obtain:

- The expected value, expected utility and list rank of all chosen (and non-chosen) lotteries.
- Each subjects risk aversion ω, calculated using the choice in the risk elicitation task.
- Each subjects search costs γ, estimated using the data from the choice rounds.

The hypotheses are divided into hypotheses regarding consumer behavior subject to different algorithms and hypotheses regarding consumers willingness-to-pay. With the former set of hypotheses I aim to document features of decisions subject to recommender algorithms in the lab setting.

4.1. Algorithms and consumer behavior

Overall, I expect that subjects make worse choices when facing the platform algorithm, than when they face the expected utility algorithm.

Hypothesis 1: *Subjects choose lotteries with lower expected utility on average when facing the platform algorithm compared to when they face the expected utility algorithm.*

Hypothesis 2: *Subjects choose lotteries with lower expected value on average when facing the platform algorithm compared to when they face the expected utility algorithm.*

Hypothesis 3: *Subjects choose lotteries that are further down the list on average when facing the platform algorithm compared to when they face the expected utility algorithm.*

Hypothesis 4: *Subjects with high search costs are more negatively effected by the platform algorithm than subjects with low search costs.*

Hypothesis 5: *Subjects that are risk neutral or close to risk-neutral are more negatively effected by the platform algorithm than subjects with high risk aversion or high preference for risk.*

4.2. Willingness-to-pay

To gauge whether subjects correctly form expectations about the different algorithms, I elicit the subjects willingness to pay for different algorithms.

Hypothesis 6: *Subjects have a positive willingness-to-pay for the expected utility algorithm.*

Hypothesis 7: *Subjects with high search costs have a higher willingness-to-pay for the expected utility algorithm than subjects with low search costs.*

Hypothesis 8: *Subjects that are risk neutral or close to risk neutral have a higher willingness-to-pay for the expected utility algorithm than subjects with high risk aversion or high preference for risk.*

5 — Analysis

To test hypotheses 1, 2, and 3, I will use a two sided t-test to test whether $y^{\text{Platform}} \neq y^{\text{Expected Utility}}$, where y is the subject mean expected utility, mean expected value, and mean lottery rank in the rounds for which the subjects face the platform algorithm and the expected utility algorithm respectively. To test hypothesis 4 and 5, I will estimate linear models of the form:

$$y_i^A = \alpha_i + \beta_1\omega + \beta_2\omega^2 + \beta_3\gamma + \mathbb{1}(\text{A = Platform}) \times [\eta + \theta_1\omega + \theta_2\omega^2 + \theta_3\gamma] + \epsilon \quad (2)$$

Where y_i^A are the outcomes expected utility, expected value, and lottery rank in the rounds with the platform and the expected utility algorithm respectively. ω_i is individual level risk aversion and γ_i is the estimated individual search cost.

To test hypothesis 7, I will use a two sided t-test to test whether $WTP \neq 0$. Finally, to test hypotheses 8 and 9, I will estimate the linear model:

$$WTP_i^A = \alpha + \beta_1\omega + \beta_2\omega^2 + \beta_3\gamma + \mathbb{1}(\text{A = Platform}) \times [\eta + \theta_1\omega + \theta_2\omega^2 + \theta_3\gamma] + \epsilon \quad (3)$$

Statistical Power Concerning hypotheses 1, 2, 3, and 6, with a sample of 300 subjects, I have power of 0.8 to detect effect sizes of 0.271 standard deviations at $\alpha = 0.05$.

References

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2018. "Incentives in Experiments: A Theoretical Analysis." *Journal of Political Economy* 126 (4): 1472–1503. 10.1086/698136, Publisher: The University of Chicago Press.

Donelly, Robert, Ayush Kanodia, and Ilya Morozov. 2023. "Welfare Effects of Personalized Rankings." *mimeo*.

Johnson, Cathleen, Aurélien Baillon, Han Bleichrodt, Zhihua Li, Dennie van Dolder, and Peter P. Wakker. 2021. "Prince: An improved method for measuring incentivized preferences." *Journal of Risk and Uncertainty* 62 (1): 1–28. 10.1007/s11166-021-09346-9.

2022. "Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA relevance)." September, http://data.europa.eu/eli/reg/2022/1925/oj/eng, Legislative Body: CONSIL, EP.

Plott, Charles R., and Kathryn Zeiler. 2005. "The Willingness to Pay-Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review* 95 (3): 530–545. 10.1257/0002828054201387.

Reimers, Imke, and Joel Waldfogel. 2023. "A Framework for Detection, Measurement, and Welfare Analysis of Platform Bias." October. 10.3386/w31766.

Resnick, Paul, and Hal R Varian. 1997. "Recommender systems." *COMMUNICATIONS OF THE ACM* 40 (3): .

Ursu, Raluca M. 2018. "The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions." *Marketing Science* 37 (4): 530–552. 10.1287/mksc.2017.1072, Publisher: INFORMS.