Dynamics of Human Trust in Recommender Systems

Jason L Harman
Department of Psychology
Louisiana State University
Baton Rouge, LA
740-707-1499
jharman@lsu.edu

John O'Donovan
Department of Computer Science
University of California
Santa Barbara, CA
805-451-9342
jod@cs.ucsb.edu

Tarek Abdelzaher
Department of Computer Science
University of Illinois at Urbana Champaign
217-265-6793
zaher@illinois.edu

Cleotilde Gonzalez
Dynamic Decision Making Laboratory
Carnegie Mellon University
5000 Forbes Ave.
412-268-6242
coty@cmu.edu

The trust that humans place on recommendations is key to the success of recommender systems. The formation and decay of trust in recommendations is a dynamic process influenced by context, human preferences, accuracy of recommendations, and the interactions of these factors. This paper describes two psychological experiments (N=400) that evaluate the evolution of trust in recommendations over time, under personalized and non-personalized recommendations by matching or not matching a participant's profile. Main findings include: Humans trust inaccurate recommendations more than they should; when recommendations are personalized, they lose trust in inaccurate recommendations faster than when recommendations are not personalized; and participants report less trust and lower overall ratings of personalized but inaccurate recommendations compared to not-personalized inaccurate recommendations. We make connections to the possible implications of these psychological findings to the design of recommender systems.

CCS CONCEPTS • Human computer interaction (HCI) • Law, social and behavioral sciences • Interaction design

Additional Keywords and Phrases: Trust, Recommendation Systems, Performance, Experimentation, Human Factors, Theory.

1. INTRODUCTION

Over the last 25 years, automated recommender systems (RS) have attempted to help users find the right information at the right time [15]. Recently, the social web has made massive amounts of user-provided content available for analysis, making the task of identifying reliable information sources increasingly difficult. In many cases, only a small window of information exists upon which a human can make a decision about whether or not to trust a recommendation and the RS that produced it. Furthermore, that window of information changes dynamically, and humans adapt their trust judgments and their preferences accordingly.

Quite recently, a number of researchers have argued that automated accuracy metrics are not enough for evaluation of RS [7, 14, 16]. A main reason is that the overall user experience with a RS should be accounted for. We believe that a key component of this overall user experience is the trust that humans place on the RS and its recommendations.

Trust in RS has been studied from the computational perspective [4, 12, 13] at the network [4], temporal [10], and algorithmic [14] levels. However, accounting for human trust in order to build RS that dynamically adjust to the human preferences and experiences is a challenge. This is largely due to the lack of research regarding how humans develop and adjust their trust in a RS. The current research is an interdisciplinary effort, bringing together behavioral and computer scientists, to collaborate on the development of recommender systems that are aware of and can adapt to the dynamics of human behavior.

This paper aims to provide generalizable scientific insight into the dynamics of human trust in RS. We use psychological experiments in simple learning paradigms to develop theoretical insights of the process of learning, choice, and trust that could inform the development of RS.

In Experiment 1, we look at the process of choosing among options of differing quality in the presence of accurate or inaccurate recommendations that are impersonal in terms of pre-defined human preferences. Results indicate that participants decrease their trust when the RS is inaccurate, yet even after extended practice with inaccurate recommendations, they continue to trust the RS more than they should objectively.

In Experiment 2, we add a richer context and personalized recommendations that match the user's profile exactly. As in Experiment 1, participants chose among options that differ in their outcome quality under accurate and inaccurate recommendations. Our results indicate that participants were able to abandon the inaccurate RS more rapidly than in Experiment 1. Furthermore, participants reported lower levels of trust and overall quality of the RS in the inaccurate condition compared to the first experiment.

We discuss implications of these findings and potential applications to the design of cognitively-aware RS.

2. BACKGROUND

Trust in recommendations systems has been approached mostly from a computational perspective. For example, early trust metrics for RS include Golbeck's metrics in social networks [4]; O'Donovan and Smyth [14] evaluated a trust model to accompany similarity scores in a collaborative filtering algorithm; and this model was extended by Liu [11] to account for temporal sequences. A common thread among these studies is that they evaluate trust in a RS using automated accuracy metrics while placing little attention to human-based trust decisions and to the role of trust from the overall human experience. In contrast, our approach highlights trust perception actions that signal trust from the human perspective. We present experiments to demonstrate human trust decisions during recommendation repeated interactions and their perception of recommendations under different accuracies of the systems.

Beyond the computational approaches above, researchers have explored how contextual elements of RS influence trust. Swearing and Sinha [16] examined interaction design for RS, focusing on user requirements and available system features. They argue that a familiar recommendation can increase trust, and that transparency of the recommendation process plays an important role. The psychology literature on conformity [1] argues that systems designed to help people make choices can have the effect of changing their subjective opinions. Herlocker [7] argues for the benefits of providing explanations to recommendations, and finds that an explanation interface can convince users to trust the system. Knijnenburg [9] and Bostandjiev [2] report similar findings using an interactive visual recommender system. Cosley et al. find that recommendations can influence users' ratings in positive or negative directions, but to a limited degree [3]. Interestingly, they report that rating shift is observed towards the predicted rating, whether it is an accurate one or not.

In the current research, we aim at advancing the theoretical basis for building RS that account for the dynamics of human trust and preferences. Particularly, we are concerned with providing some theoretical basis for building RS that dynamically adapt to the preferences and changes in human trust. This is largely a challenge in the RS community. We believe that addressing this challenge can begin with knowledge about how humans adjust their preferences and how they learn to trust or not trust RS.

In what follows, we offer results from two psychological experiments where we used simplistic paradigms of decisions from experience (DFE) [5, 6], expanded to include simple accurate or inaccurate recommendations. DFE paradigms are designed to study how humans make small daily decisions that they face repeatedly. In these well studied paradigms, participants learn from making repeated choices under conditions of uncertainty, where explicit information about outcomes and the probabilities of good outcomes from different options is not descriptively provided but rather learned from experience. The most basic DFE paradigm is presented as a money machine, where participants choose between two unlabeled buttons over several trials, receiving feedback for each choice. The buttons provide feedback based on underlying outcome distributions and over time, participants learn through experience which button provides better outcomes on average [8]. Given the research summarized above, we expected that humans would trust inaccurate recommenders more than they should, but that with extended experience, they would learn to not trust inaccurate recommendations.

3. EXPERIMENTS

3.1 Experimental Paradigm

In our experimental paradigm, participants are presented with four options (represented as buttons on the screen) that they are asked to choose from. On each of 200 trials, participants choose one of the four options and receive feedback (positive or negative) from their selected option before moving to the next trial. Positive feedback is recorded and added to a running total which determines the participant's compensation at the end of the experiment. The four options varied in the frequency with which they provided a quality outcome as: .2, .4, .6, and .8 for the four respective buttons (presentation order and labeling were randomized). The feedback probability of the four options was determined through extensive pretesting to ensure that most participants would be able to learn to discriminate between the four buttons by the end of 200 trials.

On each trial, one or more options are recommended to the participant by an accurate (recommendations always led to a good quality outcome) or inaccurate (recommendations led to a good quality outcome only half of the time) recommender. Note that our definition of accuracy represents a probability, and differs from the typical rating-based concept of accuracy in RS, e.g., from [3]. Thus, for accurate recommendations, each chosen option that was recommended on a trial produced a positive outcome (and each source not recommended produces a negative outcome); and in the inaccurate condition, recommended options produced a positive outcome half of the time and a negative outcome half of the time (the same is true for options not recommended).

The primary dependent variables in each condition are proportion of recommendations followed (as a measure of trust in the recommendations), total outcome (sum of positive feedback), and choice proportion for each of the four options (as a measure of detection of more profitable options). After participants complete 200 trials, they answer questions designed to elicit their perceptions of both the recommender and the quality of outcomes from the four options, in addition to basic questions about trust and trust in RS.

3.2 Experiment 1

Two hundred participants completed Experiment 1 on Amazon MTurk; 80 participants were female with an overall mean age of 34.7.

Experiment 1 was designed as an abstract learning and trust experiment, similar to paradigms common in cognitive science [6]. The four options were labeled A-D and the outcome feedback was composed of 1 (positive) or 0 (negative). In the experimental instructions, participants were told that the study was designed to explore how intelligence analysts collect and acquire information, and that the four options could represent sources of information such as military reports or social media that could provide information at a given time that is useful or not useful. Additionally, participants were told that a RS would highlight specific sources on each trial that could provide useful information on that trial.

3.2.1 Results

The left side of Figure 1 plots the proportion of recommendations followed on each trial for the accurate and inaccurate conditions. In the accurate recommendation condition, participant immediately chose from the recommended options and continued to choose recommended options across the 200 trials. In the inaccurate recommendation condition, participants began choosing recommended options but decreased their choice of recommended options as trials progressed, choosing recommended options about 60% of the time by the end of the experiment, indicating that participants trust inaccurate recommendations more than they objectively should (50%).

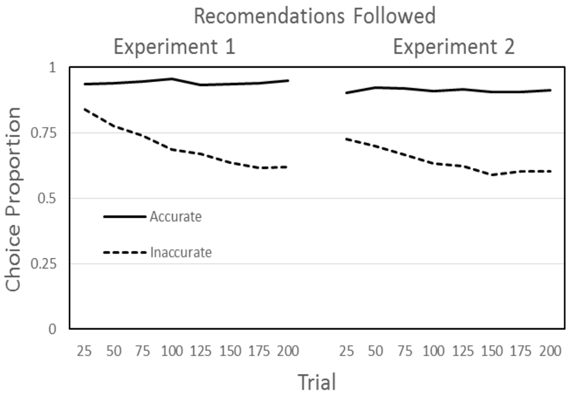

Figure 1. The proportion of choices from recommended options are plotted over time by condition (accurate recommender, inaccurate recommender) for each experiment.

Choice proportions for the four options in the accurate condition mirror the distribution of recommendations for each option (as recommendations were predominantly followed) with clear distinction between the high probability options and the low probability options. In the inaccurate condition, choice proportions between the four options (left side of Figure 2) differed across trials with participants choosing more from the best option, despite inaccurate recommendations.

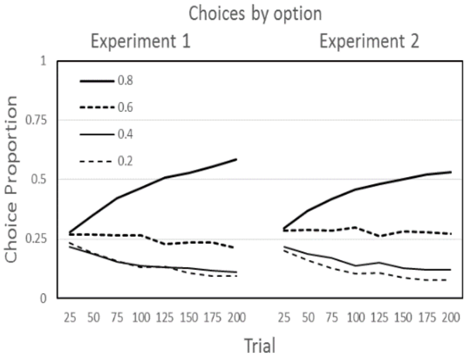

Figure 2. For the inaccurate recommender conditions only, the proportion of choices for each option (.8[best], .6, .4, .2[worst]) are plotted across time for each experiment.

3.3 Experiment 2

Experiment 2 was designed to provide personalized recommendations in a more detailed scenario. Two hundred participants completed Experiment 2 on Amazon MTurk (81 female participants, *M* age = 33.5). Participants were told that the experiment was designed to test a personalized social media app that provided recommendations on which venue to visit on a given night to have the best opportunity to meet a person they are compatible with. Before beginning the experiment, participants rated their preferences in a potential partner on three attributes with three levels in each attribute: attractiveness, education, and common interests. The participants' ratings were used to determine what feedback was considered a match (positive outcome) or not (negative outcome). The choice portion of Experiment 2 was identical in structure to Experiment 1 with two exceptions. First, the four options, instead of being labeled A-D, were labeled: Alpha Club, Beta Bar, Common Club, and Delta Bar. Second, the feedback was presented as the profile (levels of the three attributes represented by stars) of a person they meet. If the attribute levels were identical to the participant's stated preferences, it was considered a match and added to their total, which determined their monetary compensation at the end of the experiment.

3.3.1 Results

The right side of Figure 1 plots the proportion of recommendations followed in each trial for the accurate and inaccurate conditions. In the accurate recommendation condition, the proportion following recommendations mirrored that of Experiment 1. That is, people trusted and followed the recommendations almost all of the time. In the inaccurate recommendation condition, participants decreased their choice in the recommended options more over time. Compared to Experiment 1's results, the personalization with individual profiles led to participants abandoning recommendations more quickly, evident in the difference in recommendations followed in the first 25 trials (72% vs. 84% in Exp. 1). Yet, even in this case and after 200 trials, participants trusted the recommendations more than they should have (about 60% of the time by the end of the experiment).

Despite the difference in recommendations followed, the choice proportions between the four options in the inaccurate recommender condition (right side of Figure 2) is similar to the matching condition in Exp. 1, favoring the best option over time, and reducing the proportion of choices from the less favorable options over time. This is consistent with people learning the best experienced outcome from repeated trials [6].

3.4 Statistical comparisons

For each condition in both experiments, Table 1 shows the total proportion (and Standard Deviation) of the total number of choices made from recommended options, the final outcome (measure of overall performance), and the proportion of choices from the best option and the worst option. The accurate recommendations conditions did not differ from one another, and we focus our analysis on comparisons between the inaccurate conditions in Experiments 1 and 2.

The differences in behavior between the inaccurate conditions in Experiments 1 and 2 are significant for both the number of choices from recommended options (*t* (199) = 8.77, *p* < .001) and the obtained outcome (*t* (199) = 3.28, *p* < .001). There is no difference between the two groups in terms of choices from the best option 1 (*t* (199) = 0.69, *p* = .488) however the difference in choice proportion for the worst option is significant (*t* (199) = 3.18, *p* < .01).

Table 1. Mean (SD) dependent variables for each condition.

	Experiment 1		Experiment 2	
	Accurate	Inaccurate	Accurate	Inaccurate
Rec. followed	.94 (.02)	.70 (.09)	.91 (.02)	.60 (.07)
Outcome	.94 (.02)	.60 (.07)	.91 (.02)	.64 (.10)
Best Option	.47 (.06)	.46 (.10)	.45 (.06)	.44 (.27)
Worst Option	.28 (.05)	.14 (.05)	.29 (.05)	.11 (.08)

3.5 Contingent Trust Dynamics

To further explore the role of trust in choice behavior for the inaccurate conditions, we examined dynamic choice behavior in response to choosing a recommended option and receiving either a positive or negative outcome. We used two time points (trial 25 and trial 175) to explore early and late changes in trust and reactions to feedback. At each time point, we looked at participants who chose a recommended option and whether they received a positive outcome (the recommendation was correct) or whether they received a negative outcome (the recommendation was incorrect). We then calculated how many times participants chose a recommended option over the next ten trials to calculate the probability of choosing a recommended option after a successful or unsuccessful recommendation. These contingent choice dynamics are plotted in Figure 3 (the same analysis was performed using only the following trial and the following 25 trials, both of which produce the same results presented here).

Across both experiments there is a decrease in the probability of choosing a recommended option following either a successful or unsuccessful recommendation from time 1 to time 2 (Experiment 1: F (1,141) = 4.127, p < .05; Experiment 2: F (1, 96) = 71.13, p < .001). In Experiment 2, this main effect is in the presence of an interaction where the probability of choosing a recommended option after a successful recommendation is lower than after an unsuccessful recommendation at time 2 (F (1, 96) = 39.29, p < .001). This interaction did not reach significance in Experiment 1.

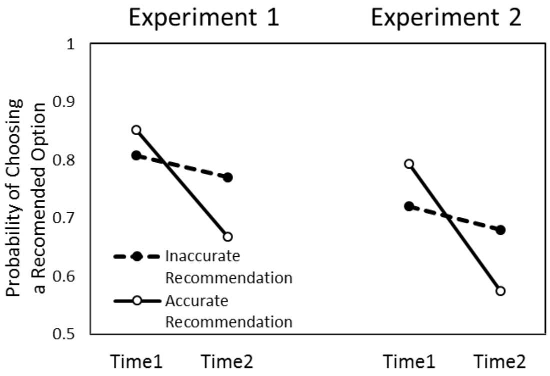

Figure 3. Probability of using the recommendation system after a correct or incorrect recommendation over time in the inaccurate recommendation condition.

3.6 Perception of the recommendations and accuracy

After the experiment, we asked participants to estimate out of ten recommendations, how many would be successful. We also asked them to rate, on a 1-5 scale, the quality of the RS and their trust in the recommendations.

Participants in the inaccurate condition of both experiments correctly estimated that the rate of accurate recommendations was about 50% (Experiment 1: M = 5.2/out of 10; SD=1.2; Experiment 2: M = 5.06, SD= 1.3). However, when asked for their opinion of the RS, participants in Experiment 2 rated it significantly worse than those in Experiment 1 (t (200) = 7.377, p < .001). Additionally, participants in the inaccurate condition of Experiment 2 reported trusting the recommendation system less than those in Experiment 1 (t (200) = 5.189, p < .001).

4. CONCLUSIONS

Several important conclusions arise from the results of these two experiments. First, humans trust recommendations, even when they are inaccurate, and they trust them more than they objectively should. This result agrees with Cosley et al.'s study [3], which found that people's opinions can be influenced based on what the RS predicts, regardless of predictive accuracy. However, in contrast to [3], we observe a difference between accurate and inaccurate conditions. Second, people lower their trust in the recommendations when they are inaccurate, but they do so more when recommendations are personalized. Third, people learn to select more accurate options even in the presence of inaccurate recommendations. And fourth, even when people are sufficiently aware of the inaccuracy of recommendations, participants feel more dissatisfied and trust the recommendations less when they are personalized.

In terms of insight for the design of real world RS, we believe that the results of these cognitive experiments are a step towards understanding when recommendation consumers are at higher risk of losing trust in a system's predictions, and conversely when their trust is high and they can be most influenced by the system's predictions.

5. ACKNOWLEDGMENTS

This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied by the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

6. REFERENCES

[1] S. E. Asch. Effects of group pressure upon the modification and distortion of judgments. In *Groups, leadership and men; research in human relations*, pages 177–190. Carnegie Press, Oxford, England, 1951.

[2] S. Bostandjiev, J. O'Donovan, and T. Höllerer. TasteWeights: A visual interactive hybrid recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, pages 35–42. ACM, September 2012.

[3] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 585–592. SIGCHI, April 2003.

[4] J. A. Golbeck. Computing and applying trust in web-based social networks. PhD thesis, University of Maryland, 2005.

[5] C. Gonzalez. The boundaries of instance-based learning theory for explaining decisions from experience. In *Progress in Brain* Research, pages 73-98. Elsevier, Amsterdam, 2013.

[6] C. Gonzalez and V. Dutt. Instance-based learning: Integrating sampling and repeated decisions from experience. *Psych. Rev.*, 118(4): 523-551, October 2011.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1): 5–53, January 2004.

[8] R. Hertwig, G. Barron, E. U. Weber, and I. Erev. Decisions from experience and the effect of rare events in risky choice. *Psych. Sci.*, 15(8): 534-539, August 2004.

[9] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, pages 43–50. ACM, September 2012.

[10] N. K. Lathia. Evaluating collaborative filtering over time. PhD thesis, University College London, 2010.

[11] D.-R. Liu, C.-H. Lai, and H. Chiu. Sequence-based trust in collaborative filtering for document recommendation. *Int. J. Hum.-Comput. Stud.*, 69(9): 587–601, August 2011.

[12] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, pages 17–24. ACM, October 2007.

[13] P. Massa and P. Avesani. Trust metrics in recommender systems. In *Computing with Social Trust*, J. Golbeck, Ed., pages 259–285. Springer, London, 2009.

[14] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 167–174. ACM, January 2005.

[15] P. Resnick and H. Varian. Recommender systems. *Commun. ACM*, 40(3): 56–58, March 1997.

[16] K. Swearing and R. Sinha. Interaction design for recommender systems. In *Designing Interactive Systems 2002*. ACM, June 2002.