# Making Sense of Recommendations

Michael Yeomans
Harvard University

Anuj K. Shah
University of Chicago, Booth School of Business

Sendhil Mullainathan
Harvard University

Jon Kleinberg
Cornell University

Harvard University
79 JFK Street
Cambridge, MA 02138
yeomans@fas.harvard.edu

**Abstract.** Computer algorithms are increasingly being used to predict people's preferences and make recommendations. These recommender systems, however, differ from prior prediction algorithms. Prior algorithms still relied on human input and expertise. Those algorithms simply improved human judgment by making it more consistent. But modern recommendation algorithms are not built on human models of judgment. These are the primary algorithms people encounter today, but we do not know how they compare to human judgment. Here, we compare computer recommender systems to human recommenders in a highly subjective domain: predicting which jokes people will find funny. We find that recommender systems outperform humans, whether strangers, friends, or family. Yet people are averse to relying on these recommender systems. This aversion partly stems from the fact that people believe the human recommendation process is easier to understand. It is not enough for recommender systems to be accurate, they must also be understood.

**Keywords:** Recommendations; Decision-Making; Machine Learning; Algorithms

**Disclosure.** None of the authors have any potential conflicts of interest to disclose in relation to this research. For each study, we report how we determined our sample size, all data exclusions, all manipulations, and all measures. The exact data and code from each study are available as Online Supplemental Material, stored anonymously on the Open Science Framework at http://goo.gl/8BjhMN

Computer algorithms can make all manner of predictions. And over the past two decades, the scope of these predictions has broadened significantly. One important trend has been the move beyond predicting objective outcomes (e.g., academic performance, criminal behavior, medical outcomes) to predicting subjective tastes (Grove et al., 2000). Most notably, recommender systems now predict people's preferences across a variety of domains, such as which movies and books people will enjoy, which products they should buy, and which restaurants they should visit (Adamovicius & Tuzhilin, 2005; Resnick & Varian, 1997). These recommender systems help people in many markets find the items they want by reducing search costs (Ansari, Essegaier, & Kohli, 2000; Brynjolfsson, Hu, & Simester, 2011). And recommender systems can have a significant impact on firm revenues (Bodapati, 2008; De, Hu & Rahman, 2010). In some cases, a company will even tie its reputation to the strength of its recommendations, as with the Netflix Prize for building a better recommender system (Bell & Koren, 2007; Gomez-Uribe & Hunt, 2016).

Of course, people have long relied on recommendations to inform their choices, but these recommendations have typically come from other people (Berger, 2014; Bonaccio & Dalal, 2006). Whether deciding where to eat, what movie to watch, or even whom to date, we rely on the opinions of friends, family, and even strangers on the internet. And people trust other people to provide good recommendations—83% of people trust recommendations from friends and family; 66% trust online opinions of strangers (Nielsen, 2015). But given that algorithmic recommendations now play a larger role in curating our experiences, it seems natural to ask how well recommender systems perform. Specifically, how do recommender systems compare to human recommenders?

This question builds on a large body of work comparing human and algorithmic judgment. But it diverges from prior work in theoretically and practically important ways. Prior algorithms capitalized on the psychological insight that human judgment can be improved by making it more consistent

(Dawes, 1979; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954). Models of judgment assume that people identify relevant features or cues and use them to make judgments (Brunswik, 1952; Karelaia & Hogarth, 2008). Early research discovered that algorithms, when given these features, produce more accurate judgments than do human judges who are using the same features. That seminal work posits a key role for humans in constructing machine intelligence: In each domain (e.g., predicting academic performance or medical outcomes) humans identify the relevant features (e.g., student's GPA, patient's age) for algorithms to use. As Dawes noted, "The linear model cannot replace the expert in deciding such things as 'what to look for,' but it is precisely this knowledge…that is the special expertise people have" (1979, p. 573).

But today's algorithms are built on different principles. Consider commonly used algorithms for recommendations about movies to watch (or any other preference domain). A Dawesian recommendation algorithm would be based on human models of what drives preferences. For example, to recommend a movie, it might use features such as genre, actors, and time-period (or any other features a person deems relevant). Modern recommender systems, however, need no such model. Instead, they can rely solely on a database of people's ratings of past experiences. "Collaborative filtering" algorithms, for instance, simply learn which experiences are *statistically* similar (Breese, Heckerman & Kadie, 1998; Koren & Bell, 2011). To recommend a movie, these systems would start with a database of movies that have been rated by multiple people. Movies are said to be "similar" if they have correlated ratings (the simplest similarity score would be to compute the Pearson correlation of ratings for two movies across all people who rated both movies). The system would then recommend a movie whose ratings correlate strongly with the movies that a person rated highly in the past.

Modern day recommender systems take a fundamentally different approach to prediction than humans or Dawesian algorithms do. Recommender systems rely on ratings data, but they do not know

what is being rated. Humans and Dawesian algorithms rely on content. In fact, when given a choice between basing their predictions on content (e.g., a description of an experience) or mere ratings data (how much someone enjoyed that experience), people prefer content (Gilbert, Killingsworth, Eyre, & Wilson, 2009). Put simply, recommender systems make recommendations for movies they will never see, books they will never read, and restaurants they will never visit. But for humans, content or experience is essential to recommendations. And it implicitly guides the Dawesian algorithms that are based on human judgment.

As a result, human expertise enters very differently into today's recommender systems. Humans program Dawesian algorithms and build in knowledge of what features of a domain to "look for" in making predictions. With recommender systems, humans program a general purpose statistical algorithm. They do not build in any insights about what characteristics make a movie or book or restaurant enjoyable. Tellingly, recommender systems need little human expertise to adapt to different domains. The exact same statistical procedure is applied to any database of ratings, whether they are of restaurants, books, or cars. Humans also inform the algorithms through the data on what they have liked. But in providing these data, humans do not tell the algorithm "what to look for."

Thus, comparing human recommenders to recommender systems presents a different test than prior comparisons of human and algorithmic predictions. When humans were compared to Dawesian algorithms, they were essentially compared to routinized versions of a human process. Comparing humans to recommender systems presents a test of whether an algorithm can outperform humans even when the algorithm is not modeled after human judgment.

Unlike Dawesian algorithms, recommender systems may fare poorly because they do not benefit from human expertise. They may only be popular because they are convenient and cheaper to scale than human recommendations. In fact people seem to have many advantages, particularly for

predicting subjective preferences. They have direct experience with what they are recommending and often know details about the lives of the people who will receive their recommendations. By contrast, recommender systems operate in the dark. They have limited information about the unique tastes of the recipient, and no direct knowledge of what they are recommending (e.g., they do not consider the plot of a movie or the text of a book). They only know *what* people like, not *why* people like it. On the other hand, this could give recommender systems an advantage. A growing body of research suggests that people can often better predict their own enjoyment of an item if they know how much someone else liked that item than if they know more details about the item (Eggleston, Wilson, Lee, & Gilbert, 2015; Gilbert et al., 2009).

In this paper, we present the first rigorous test of whether recommender systems or humans more accurately predict preferences. In our studies, we acquire human and algorithmic recommendations about which jokes people will like. We then compare these recommendations to which jokes people actually liked. Initial research in this area has not provided a definitive answer on which recommender is superior because no study has measured whether people actually enjoy the recommended items (Krishnan et al., 2008; Sharma & Cosley, 2015; Sinha & Swearingen, 2001).

We chose joke recommendations as our test domain for several reasons. First, it takes very little time to read and respond to jokes. This makes it an ideal domain for having people experience novel instances and make several ratings and recommendations in a relatively short time frame. Doing so with other preference domains (e.g., movies, books, restaurants, dating partners) would be less practical. But there are also reasons to believe that humor is fairly representative of other preference domains. As with other matters of taste, humor is highly subjective. Moreover, it is something with which most people have significant experience, but one does not need to be a connoisseur to appreciate it. And, as in other domains, there are genres of humor and heterogeneity in people's preferences for

those genres (some people appreciate political humor, others enjoy ribald jokes, etc.).

For our studies, we created a simple recommender system based solely on principles of collaborative filtering. More sophisticated systems could combine humans' domain expertise with the logic of collaborative filtering. But to have the crispest test, our system does not do this (it is also worth noting that none of the authors have special expertise in being funny). The algorithm has no model of what features make a joke funny, nor does it parse the language of the jokes. It simply relies on correlations between ratings of jokes.

Our first finding (Studies 1A-1B) shows that, despite clear disadvantages, recommender systems outperform human recommenders, even for a highly subjective domain that might be uniquely human. They are better than humans at picking jokes that people find funny. This is true regardless of whether the human recommenders are strangers, friends, family, or significant others.

However, our second result highlights a familiar tension: People are averse to using recommender systems. We find that when people are making recommendations for others, they are reluctant to use input from a recommender system that would have improved their recommendations (Study 2). Moreover, we find that people would rather receive recommendations from a human than from a recommender system (Study 3). This echoes decades of research showing that people are averse to relying on algorithms (for a review, see Dietvorst, Simmons, & Massey, 2015). With Dawesian algorithms, the primary driver of aversion is algorithmic errors. This might also explain some of the aversion to recommender systems, but our final two studies suggest that there is an additional factor beyond aversion to error.

Prior research suggests that people want recommender systems to provide recommendations that they can make sense of (Herlocker et al., 2000; McNee, Reidl & Konstan, 2006). But because these systems are not modeled after human judgment, the recommendation process (and therefore the

recommendations) might be particularly difficult to understand. Indeed we find that people think recommendations are easier to understand when they come from a human instead of an algorithm (Study 4). However, these feelings of subjective understanding are often quite malleable (Keil, 2003; Rozenblit & Keil, 2002; Tintarev & Masthoff, 2011), which makes it possible to reduce aversion to algorithms more easily. We find that people are less averse to recommender systems when you simply explain how they work (Study 5). Thus, it is not enough for algorithms to be more accurate, they also need to be understood.

This paper therefore makes three contributions. First, we test how human and algorithmic judgment compare when predicting *preferences*, which prior research has overlooked, but which is a dominant application of algorithms today. Second, the algorithms that we study here are not modeled after human intelligence. Indeed, the algorithms studied here are not given access to the features humans are using to make judgments. Finally, we show that aversion to algorithms does not merely stem from concerns about algorithmic accuracy. Instead, it also depends on whether people can understand the algorithms.

For all studies, sample sizes were set a priori and analyses were not conducted until all data were collected. A priori, we also determined five reasons for excluding participants: (1) They did not pass the initial attention check (see Appendix A), (2) they did not complete the study, (3) they did not follow instructions, (4) they failed a manipulation check (see Appendix B), or (5) they rated all jokes as equally funny. The full set of all measures from every study (including exclusion criteria) are described in the Supplemental Material, and all data and analyses are posted on http://goo.gl/8BjhMN

## STUDY 1A

**Methods**

One hundred fifty participants (75 pairs) were recruited from the Museum of Science and Industry in Hyde Park, Chicago. Twenty-eight participants (14 pairs) were dropped due to incomplete responses or not following instructions, leaving 122 participants (61 pairs). All pairs had come to the museum together, and most pairs knew each other very well (e.g., family, friends, partners).

Every participant both received recommendations (i.e., was a "target") and made recommendations (i.e., was a "recommender"). Participants were seated at separate computer terminals where they could not see or hear each other. First, participants completed the ratings phase of the experiment. They saw 12 jokes (taken from Goldberg et al., 2001) presented in a random order; all participants saw the same jokes. Participants rated each joke on a scale from -10 (not funny at all) to +10 (extremely funny).

Next, participants completed the recommendation phase of the experiment (see Appendix C for stimuli). Participants switched computer terminals, where they saw four of the jokes (the "sample set"), randomly selected from the full set. They were also shown their partner's ratings for those sample jokes. They then predicted their partner's ratings for the remaining eight jokes (the "test set"). Thus, we had targets' *actual* ratings of the test jokes, and recommenders' *predictions* about the targets' ratings of the test jokes (note: recommenders never saw targets' ratings of the test jokes).

Our algorithm runs a series of regressions to determine which sample jokes are most similar to each test joke (Sarwar, Karypis, Konstan & Riedl, 2001). For each of the eight test jokes, it runs a separate linear regression where the dependent variable is a user's rating of the test joke and the independent variables are their ratings of the four sample jokes. This regression assigns weights to each of the four sample jokes. The sample joke ratings can then be used to predict the user's ratings for the test joke.

Of course, in forming the predictions for a particular person, we would not want to use the person's own data in these regressions. To avoid this problem, we use "leave-one-out cross-validation": When forming predictions for a particular user, the model is trained on data from *all other* users. This ensures that we are not making predictions for users who were used to train the model. Thus, the data are recycled so that every subject is used for both testing and training. Both human and machine recommenders made predictions using the same -10 to +10 scale that participants used to rate the jokes. Prediction error was defined as the squared difference between each prediction and its true value (i.e., the target's actual rating of the joke), where larger errors indicate less accurate predictions. We compared the accuracy of predictions from human recommenders and our recommender system.

**Results**

Our recommender system made predictions that were "yoked" to human recommenders' predictions. That is, for a given target, human recommenders made eight predictions based on the four sample joke ratings. Our recommender system did this as well. We then computed the average error across these eight predictions, and compared the average error for human recommendations to the average error for machine recommendations. The recommender system was more accurate (*RMSE* = 4.281, bootstrapped *95% CI* = [4.126, 4.448])) than human recommenders (*RMSE* = 5.586, bootstrapped *95% CI* = [5.360, 5.841]), $t(121) = 6.90$, $P < .001$.

One concern might be that the human recommenders simply were not trying very hard at the task. But we do see evidence that human recommenders were trying to be accurate. For instance, prior research (Eggleston et al., 2015; Gilbert et al., 2009; Hoch, 1987) suggests that people can make remarkably accurate recommendations simply by acting as "surrogates" for the target (i.e., the person receiving the recommendation). In this case, surrogation would mean that recommenders predicted that

the target would give a joke the same rating that the recommender did. In our study, human recommenders outperformed mere surrogation (*RMSE* = 6.495, bootstrapped *95% CI* = [6.291, 6.699]), $t(121) = 5.81$, $P < .001$. The fact that our participants outperformed this benchmark suggests that they were invested in the task. But they could not match the performance of the recommender system.

To our knowledge, this is the first experiment that compares people's actual enjoyment of items recommended by machines and humans. We find that recommender systems are more accurate predictors. In this design, our recommender system even outperformed people who know each other well. But this study may have disadvantaged participants. Perhaps participants' judgments were *clouded* by knowing each other well (Davis, Hoch & Ragsdale, 1986; Lerouge & Warlop, 2006). If so, then a fairer test may be to have humans make recommendations for strangers. Participants were also disadvantaged because the recommender system "sees" more ratings in the database. As a result, the recommender system can calibrate its use of the scale better than human recommenders can. The next study addresses both of these issues.

## STUDY 1B

### Methods

Five hundred eighty-one participants from Amazon.com's Mechanical Turk (MTurk) platform completed our study. Thirty-four failed the attention check, and three gave the same rating to every joke, leaving 544 participants for the analyses. Participants served only as recommenders, not targets. The targets were instead drawn from a pool of previous participants who had rated the same 30 jokes in other experiments.

Each participant in our study made recommendations for five targets randomly drawn from the pool. For each target, participants were shown the text of four sample jokes, along with the target's ratings of those jokes. Then, for each target, participants predicted the ratings of two test jokes (10 total

predictions). Thus, participants saw all 30 jokes exactly once, but the order of the jokes (and whether a joke was given as a sample or test joke) was randomly determined for each participant. Accuracy was incentivized by giving a $20 bonus to the most accurate participant. At the end of the study, participants rated each joke.

There were two conditions in this study. Half of the participants were assigned to the "base rate" condition, where they were told the mean rating for each test joke. That is, when predicting a target's rating for a joke, they were shown the average rating for that joke across all other targets in the database. This would help participants calibrate their sense of how to use the scale for each joke. The other half of participants were assigned to the "no information" condition, which was essentially identical to the paradigm used in Study 1A. Machine recommendations were produced using the same method as in Study 1A (i.e., leave-one-out cross-validation was used to train the recommender system).

**Results**

Once again, machine recommenders outperformed human recommenders. Specifically, the recommender system was more accurate (*RMSE* = 4.645, bootstrapped *95% CI* = [4.576, 4.715]) than humans in the "no information" condition (*RMSE* = 6.087, bootstrapped *95% CI* = [5.932, 6.255]), $t(247) = 15.22$, $P < .001$, as well as humans in the "base rate" condition (*RMSE* = 5.873, bootstrapped *95% CI* = [5.726, 5.993]), $t(271) = 12.87$, $P < .001$. Moreover, humans recommenders were only slightly more accurate when they were given base rate information $t(518) = 1.39$, $P = .166$. This suggests that the recommender system outperforms humans even when they have information about how people use the scale.

**Robustness and discussion**

Taken together, Studies 1A and 1B clearly suggest that recommender systems can outperform human recommenders, even for a highly subjective domain, and regardless of whether the

recommendations are made for strangers or for close others. Remarkably, these recommender systems are able to outperform humans even without the "special expertise" from humans that informed more traditional Dawesian prediction algorithms.

However, there may be two lingering concerns about this finding. First, did we select an appropriate domain for comparing human recommenders and recommender systems? One worry might be that people simply do not have very heterogeneous preferences for jokes. If people had homogenous preferences in this domain, then our result would be little more than a repackaged wisdom-of-crowds effect (Clemen, 1989; Galton, 1907). Humans might excel at detecting idiosyncratic preferences, but this domain would prevent them from being able to do so. Meanwhile, our recommender system would excel because of the statistical advantages of averaging, but not necessarily because collaborative filtering allowed it to tailor its recommendations to people's idiosyncratic preferences (Hoch, 1987; Muller-Trede et al., 2017). Put simply, if we selected a domain with insufficient heterogeneity, then our results would not tell us whether collaborative filtering outperformed humans, and it would not have given humans a chance to excel.

To test this possibility, we compared the recommender systems' predictions to a benchmark of simple averaging. Specifically, each time the recommender system predicted a target's rating for a joke, we compared that predicted rating to the average rating for the joke across all participants (except the target) in the database. In Study 1A, the recommender system (*RMSE* = 4.281, bootstrapped *95% CI* = [4.117, 4.439]) outperformed the simple average (*RMSE* = 4.606, bootstrapped *95% CI* = [4.467, 4.762]; $t(975)=3.8$, $P<.001$). This was also true for Study 1B (machine: *RMSE* = 4.645, bootstrapped *95% CI* = [4.571, 4.716]; average: *RMSE* = 4.966, bootstrapped *95% CI* = [4.921, 5.006]; $t(5199)=12.7$, $P<.001$). These results suggest that there is reasonable heterogeneity across people's preferences in this domain, and the recommender system is able to outperform human recommenders by detecting

these idiosyncrasies.

A second concern might be that we disadvantaged humans by asking them to predict absolute ratings instead of making comparative judgments. Perhaps it is difficult for people to identify the "funniness" of a joke on a scale, while it would be easier for people to simply state which of two jokes someone would like more. We can re-analyze our data from Studies 1A and 1B to test this possibility.

For Study 1A, each recommender made eight recommendations. This allows us to compute 28 pairwise comparisons: For each pair, we would know which joke the target *actually* rated higher, and which joke the recommenders *predicted* to be rated higher. If a recommender gave a higher rating to the item in the pair that the target actually rated higher, then this was scored as a correct response (ties were counted as half-correct). Each recommender's accuracy was calculated as their average over all 28 pairwise comparisons. This pairwise analysis ensures that humans are not punished for miscalibrated absolute judgments of funniness. Once again, the recommender system outperformed ($M$ = 61.1%, *95% CI* = [58.7%, 63.4%]) human recommenders ($M$ = 56.8%, *95% CI* = [54.2%, 59.5%]), $t(121) = 2.65$, $P = .009$. For Study 1B, each recommender made two recommendations for each of five targets. This allows us to compute five pairwise comparisons per recommender. Once again, the recommender system ($M$ = 60.4%, *95% CI* = [58.5%, 62.3%]) was more accurate than the human judges ($M$ = 54.8%, *95% CI* = [52.8%, 56.7%]), $t(519) = 4.91$, $P < .001$.

Finally, in another study (see Appendix D), we asked participants to directly make pairwise comparisons when producing their recommendations. Even then, machine recommenders ($M$ = 62.9%, *95% CI* = [59.8%, 66.1%]) were more accurate than human recommenders ($M$ = 56.6%, *95% CI* = [53.6%, 59.7%]), $t(196) = 3.15$, $P = .002$. These results suggest that machines did not outperform humans merely due to an artifact of the recommendation procedure.

These initial studies provide the first rigorous evidence that recommender systems can

outperform human recommenders. And they do so without relying on a human to specify which features of jokes are most predictive of how funny they will be. They do this without forming a model of how humans make recommendations. But because these recommender systems do not follow a human model of prediction, people might be averse to the idea of relying on these recommenders too heavily. In our remaining studies, we continue to develop evidence that recommender systems outperform human recommenders, but our focus now shifts to a related question: Despite the superiority of machine recommenders, are people averse to using them?

## STUDY 2

### *Methods*

We recruited 232 participants (116 pairs) from the Museum of Science and Industry; 22 participants (11 pairs) were dropped due to incomplete responses or not following directions, leaving 210 participants (105 pairs).

The procedure closely paralleled Study 1A, with a few differences. Participants were assigned to one of four conditions in a 2x2 between-subjects design. The first factor was whether participants were given machine recommendations to guide their own recommendations. In the "with machine" condition, participants were told about the database of joke ratings and were given an explanation of collaborative filtering. During the recommendation phase of the experiment, these participants were shown the machine's predicted rating for each test joke. Participants were told that these predicted ratings could be used to inform their own predictions, or they could ignore them if they wished. In the "without machine" condition, participants were not given the machine's predicted rating (or told about collaborative filtering).

To generate the machine's predictions during the current study, we needed to build the

recommender system prior to conducting the study. This recommender system was developed using the data and procedures from Study 1A.

We were unsure whether people would rely on the machine predictions more when making recommendations for strangers or people they know. Accordingly, the second factor in our experiment manipulated the target of the recommendation. Participants in the "known" condition made recommendations for the other person in the pair. Participants in the "stranger" condition made recommendations for someone selected at random, whom they did not know (i.e., they were shown sample ratings for a stranger whose data we already had and were told they were predicting that stranger's ratings). Both factors were randomized at the pair level (i.e., people recruited together were always in the same condition).

***Results***

Regarding accuracy, recommender systems (*RMSE* = 4.273, bootstrapped *95% CI* = [4.147, 4.402]) once again outperformed humans (*RMSE* = 5.387, bootstrapped *95% CI* = [5.231, 5.558], t(209) = 10.06, $P < .001$). And the humans did not perform any better for close others (*RMSE* = 5.386, bootstrapped *95% CI* = [5.190, 5.614]) or for strangers (*RMSE* = 5.387, bootstrapped *95% CI* = [5.167, 5.613]), $t(208) = 0.25$, $P = .802$.

Despite the fact that machines were more accurate than humans, humans showed some aversion to using the machine recommendations. Humans did improve somewhat when given the machine predictions (*RMSE* = 5.056, bootstrapped *95% CI* = [4.858, 5.272]) compared to those without it (*RMSE* = 5.692, bootstrapped *95% CI* = [5.465, 5.943]), $t(208) = 2.42$, $P = .017$. But the humans with the recommender system still performed worse than the recommender system on its own (*RMSE* = 4.110, bootstrapped *95% CI* = [3.948, 4.293], $t(103) = 6.61$, $P < .001$). These data suggest that people

do not completely ignore the machine recommendation. But they are averse enough to using the machine recommendation that they did not perform as well as they could have. These results echo prior research which has shown that people are reluctant to use many kinds of judgment or decision aids (Bar-Hillel, 1980; Dietvorst et al., 2015; Larrick & Soll, 2006; Mannes, 2009; Yaniv, 2004).

However, this study may not provide the most direct or important test of whether people are averse to machine recommendations. People rarely use recommender systems to help them *make* recommendations. Instead, people most often interact with recommender systems when they are *receiving* recommendations. Are people averse to *receiving* machine recommendations for themselves?

## STUDY 3

If people are averse to machine recommendations, then this could be due to two factors. First, machine recommenders select different content (i.e., which jokes they recommend). Second, machine recommenders use a different recommendation process than do humans. We expect that the second factor more strongly shapes people's aversion to relying on recommender systems. In this study, we disentangle these two factors by manipulating the *actual* source of recommendations (which changes the content and process) and the *perceived* source (which holds content constant).

### *Methods*

All participants in this study were targets, not recommenders. They received recommendations from either another person or from our recommender system, based on how participants rated three sample jokes.

**Developing human and machine recommendations.** Because it would be difficult to acquire human recommendations in real time, we developed a method to collect the recommendations in

advance and match them to our participants *ex post* based on participants' ratings of the three sample jokes. We rounded participants' sample ratings to the nearest 2.5-point marking on the scale, which meant that each joke rating would be rounded to one of nine scores (-10, -7.5, -5, -2.5, 0, 2.5, 5, 7.5, and 10). With three jokes in the sample set, there were $9^3$=729 possible permutations of sample joke ratings.

A separate sample of 253 MTurk participants provided the human recommendations. These recommenders were shown these ratings profiles along with the sample jokes (e.g., Sample Joke 1: 2.5, Sample Joke 2: -5.0, Sample Joke 3: 7.5). Recommenders then picked three test jokes (from a menu of ten) that they thought someone with those ratings would like most. Each recommender made three sets of recommendations. All recommendations were pooled together into a database. This database made it possible to have a human recommendation ready for every participant, because it contained recommendations for the 729 possible permutations of sample ratings that participants could produce.

Of course, recommender systems would have an unfair advantage if they used participants' precise ratings while human recommendations were based on rounded ratings. To address this concern, the algorithm also used the same rounded ratings to make predictions.

**Current study.** Nine hundred ninety-six participants from MTurk completed our study; 104 participants failed the manipulation check and 6 participants predicted the same rating for every joke, leaving 886 participants for the analyses.

Participants were randomly assigned to one of four conditions in a 2x2 between-subjects design. The first factor was the actual recommender (human or recommender system) and the second factor was the perceived recommender. Participants in the perceived-human recommender conditions were told that they were paired with another user online, although this was not true since the recommendations were collected in advance, as described above. Participants in the machine condition

were told that the recommender system would use a "database of thousands of people" to find others with a "similar sense of humor" based on the sample jokes, though we did not explain the details of the algorithms involved.

Participants first rated three sample jokes and ten test jokes. They then waited 20 seconds and were shown the three jokes from the test set that the recommender thought they would like most. After seeing these jokes, participants evaluated their recommender across three questions: (1) "How good do you think the recommender was at choosing jokes you would enjoy?" (2) "How well do you think the recommender knew your sense of humor?" and (3) "How much would you want to read more jokes that the recommender chose for you?" All responses were on a 7-point scale.

Finally, as a comprehension check, participants were asked who made the recommendations in a multiple choice question.

***Results***

**Accuracy.** For each participant, we can compare the participant's average rating of the three jokes from the test set that a human recommender selected to the participant's average rating of the three jokes that the recommender system selected. This within-subjects comparison once again shows that the recommender system picked jokes that participants found funnier ($M = 3.03$, *95% CI* = [2.79, 3.27]) than did human recommenders ($M = 2.74$, *95% CI* = [2.50, 2.98]), $t(885) = 3.01$, $P = .003$.

**Aversion to recommender systems.** Next, we compared how participants rated the *recommenders.* Because there was high internal consistency among the three evaluation questions (Cronbach's $\alpha = 0.95$), we standardized and combined responses into a single "preference index." A 2x2 ANOVA revealed a significant main effect of perceived recommender on these evaluations. Participants rated the recommender more highly when they believed it was human ($M = 0.07$, $SD =$

1.01) than when they believed it was a machine ($M$ = -0.07, $SD$ = 0.98), $F(1, 882) = 4.6$, $P = .032$. However, there was not a significant effect of the actual recommender (human: $M$ = -0.03, $SD$ = 1.02, machine: $M$ = 0.02, $SD$ = 0.99; $F(1, 882) = 0.6$, $P = .432$), nor a significant interaction, $F(1, 882) = 1.8$, $P = .178$.

These results demonstrate another dimension of aversion to recommender systems. Not only are people reluctant to use recommender systems when making recommendations for others (as in Study 2), but they are also averse to recommender systems when receiving recommendations. Importantly, this aversion does not stem from the different *content* of what the machines recommend. Instead, people were averse to recommendations that simply *seemed* to come from recommender systems.

Interestingly, people do prefer more accurate recommendations. Accuracy was strongly correlated with the preference index ($r = 0.35$, $t(884) = 11.1$, $P < .001$). Nevertheless recommender systems were judged more harshly. We benchmarked the effects of actual accuracy and perceived recommendation source in a multiple regression model, which included both variables as predictors. Based on this model, we estimate that the implicit penalty against the machine was equivalent to a difference in accuracy of 0.38 standard deviations.

These findings reveal an interesting pattern—although people like the machine's *recommendations* more, they like human *recommenders* more than the recommender system. Why might this be? Perhaps it is due to differences in how people perceive the human versus machine recommendation process. It is hard for people to understand how recommender systems operate (Herlocker, Konstan & Riedl, 2000; Tintarev & Masthoff, 2011), perhaps all the more so because they do not follow a human model of judgment. It is possible that people are averse to recommender systems because it seems harder to understand how machines make recommendations than how humans make recommendations. In the next study, we test whether (a lack of) subjective understanding

of the recommendation process predicts aversion to using recommender systems.

## STUDY 4

### *Methods*

One thousand ten participants from MTurk completed our study; 107 failed the manipulation check and 4 gave the same rating to every joke, leaving 899 participants for the analyses.

The study was identical to Study 3, with two exceptions. First, participants were asked to rate how easy it was to understand the recommendation process by stating their agreement with two statements: "I could understand why the recommender thought I would like those jokes" and "It is hard for me to explain how the recommender chose those jokes" (reverse-coded). For both questions, participants responded on a scale ranging from -3 to +3, anchored at "strongly agree" to "strongly disagree", with the 0 point labelled "neutral". The order of these two questions was counterbalanced.

Second, to assess aversion, participants indicated whether they would rather receive additional recommendations from humans or from the recommender system. Participants imagined that they would receive additional recommendations from either "an algorithm [that] would search through a database of thousands of people to find jokes liked by those who had the most similar sense of humor to your own" or from "another person [that] would then choose some jokes that they thought you would like."

### *Results*

Accuracy was calculated as in Study 3. Recommender systems were once again more accurate ($M = 3.13$, *95% CI* = [2.90, 3.36]) than human recommenders ($M = 2.71$, *95% CI* = [2.46, 2.95]), $t(885) = 3.01$, $P = .003$.

To assess the relationship between aversion and subjective understanding, we collapse our

analyses across the *actual* recommender, to focus on the effective of the *perceived* recommender. This way, the actual jokes being recommended are held constant and the only thing that varies is the perceived process by which recommendations are made.

When participants were asked which recommender they would choose, most participants (74.1%) wanted to switch recommenders. Critically, more participants chose to switch when they started with a machine recommender ($M$ = 79.5%, *95% CI* = [75.8%, 83.3%]) than when they started with a human recommender ($M$ = 68.8%, *95% CI* = [64.6%, 73.1%]), $\chi^2(1, N = 899) = 12.84$, $P < .001$. Put simply, a majority of participants preferred human recommenders ($M$ = 54.8%, *95% CI* = [51.6%, 58.1%]), $\chi^2(1, N = 899) = 8.42$, $P = .004$.

The subjective understanding ratings were combined in a single index (Cronbach's $\alpha = 0.82$). Participants rated human recommenders as easier to understand ($M$ = 0.07, *95% CI* = [-0.02, 0.16]) than machine recommenders ($M$ = -0.07, *95% CI* = [-0.17, 0.03]), $t(897) = 2.07$, $P = .038$). And these beliefs were strongly related to participants' preferences for recommenders. Across all conditions, participants were more likely to stick with their assigned recommender if they thought the recommender was easier to understand (logistic regression, $\beta = 0.60$, $SE = 0.09$, $z(897) = 7.01$, $P < .001$). And this relationship was attenuated when participants thought their recommender was human ($\beta = 0.43$, $SE = 0.11$, $z(457) = 3.83$, $P < .001$; interaction term: $\beta = -0.39$, $SE = 0.18$, $z(895) = 2.19$, $P = .028$). This suggests that people are averse to using recommender systems because it seems harder to understand how machines make predictions than how humans do.

These results put our earlier findings into clearer focus. When participants thought the recommendations had come from a human, they were able to make sense of why someone might have chosen them. But when they thought the recommendations had been generated by a machine, those very same recommendations were perceived as inscrutable. These results suggest that people are less

willing to accept recommenders when they do not understand *how* they make recommendations. Would making machine recommendations easier to understand increase how much people like those recommenders? The final study addresses this possibility.

## STUDY 5

### *Methods*

One thousand and fourteen participants from MTurk completed our study. 24 participants failed the manipulation check and 4 participants gave the same rating to every joke, leaving 986 participants for the analyses.

The study was identical to Study 4, with four changes. First, participants only rated three sample jokes and then rated the three recommended jokes chosen by the recommender system. Second, all recommendations were generated by a recommender system that used the exact (i.e. un-rounded) sample joke ratings from each participant as inputs, as in Study 2. Third, the dependent measures consisted of the subjective understanding questions from Study 4, and the preference questions from Study 3. The order of these two sets of questions were counterbalanced across participants.

Finally, the most substantive change was a manipulation of how the recommender system was explained. Some participants received a *sparse* explanation. During the introduction to the study participants were told, "…we are going to feed your ratings into a computer algorithm, which will recommend some other jokes that you might also like." Other participants received a *rich* explanation, where they were also told to "Think of the algorithm as a tool that can poll thousands of people and ask them how much they like different jokes. This way, the algorithm can learn which jokes are the most popular overall, and which jokes appeal to people with a certain sense of humor. Using the database ratings, the algorithm will search for new jokes that are similar to the ones you liked, and dissimilar to the ones you did not like." The rich condition also repeated these details after the participants rated the

sample jokes when they were waiting for their recommendations, and again when the recommended jokes were shown (see Appendix E for exact stimuli).

### *Results*

Participants in the *rich* explanation condition rated the recommender system as easier to understand ($M = 0.09$, *95% CI* = [0.01, 0.18]) than participants in the *sparse* condition ($M = -0.09$, *95% CI* = [-0.18, 0.00]), $t(984) = 2.93$, $P = .003$). This confirmed that our manipulation had its intended effect. Turning to the preference questions, participants in the *rich* condition showed greater preference for the recommender system ($M = 0.07$, *95% CI* = [-0.02, 0.16]) than participants in the *sparse* condition ($M = -0.07$, *95% CI* = [-0.16, 0.02], $t(984) = 2.20$, $P = .028$). This effect was significantly mediated by subjective understanding (bootstrapped indirect effect: $M = 0.104$, *95% CI* = [0.034, 0.175], $P = .002$). In other words, rich explanations of the recommender system increased participants' understanding of the recommendation process, and this increased willingness to use a recommender system.

## GENERAL DISCUSSION

The ubiquity of recommender systems raises a familiar question: How do algorithmic predictions compare to human judgment? But recommender systems differ from previous algorithms in that they are not informed by human models of judgment. Comparing them to human judges actually raises a more novel question: Can we improve on human judgment without the special expertise of humans? Our results suggest the answer is yes. Recommender systems can outperform human recommenders, even when those humans are making recommendations for friends and family.

Of course, recommender systems cannot work without human input. They depend on people's ratings of items and experiences. But this is different from the special expertise that informed Dawesian

algorithms. For example, imagine an algorithm that predicted which kinds of flowers bees liked. To do so, it could use a database of flowers and individual bees, with data on how often each bee visited each kind of flower. This algorithm could recommend flowers for bees using the same process behind joke recommendations for humans. But neither the bees nor the humans directly inform the algorithm. In fact, it is not clear that a human (or bee) could even identify what latent features the algorithm is implicitly picking up on in the ratings data.

Still, as with previous algorithms, people are averse to relying on recommender systems. People ignore input from these systems and they prefer to receive recommendations from humans. This is a familiar tension: Recommender systems are more accurate than humans, but people prefer to receive recommendations from humans. Previous research identifies one reason why people are averse to using algorithms, namely that people are concerned about algorithmic errors (Dietvorst et al., 2015). But the aversion to recommender systems appears to run deeper. Beyond concerns about accuracy, people seem averse to recommender systems because they do not understand the recommendation process. People believe that human recommenders are easier to understand.

We should emphasize that our studies only tell us that people *subjectively* feel like they understand human recommendations better than machine recommendations. Of course, these subjective impressions need not align with reality. People might be overconfident in their understanding of how humans make recommendations. And they may not truly understand the factors that influence these subjective impressions (Nisbett & Wilson, 1977; Rozenblit & Keil, 2002). Nevertheless, people seem more comfortable with human recommenders, in part, because of these subjective feelings of understanding.

This then raises a pressing question for future research. What factors influence *algorithmic sensemaking*? This has typically been a secondary question (and our work does not offer a final answer to this question either). Instead, researchers often focus on how to engineer more accurate algorithms.

The "Netflix Challenge," for example, offered $1 million to researchers who could improve prediction accuracy by just 10% (Bell & Koren, 2007). But people are especially wary to rely on algorithms that make recommendations about subjective preferences, and increasing accuracy may not be sufficient to overcome this (Logg, 2017). In some sense, if the next "Netflix Challenge" focused on facilitating algorithmic sensemaking, it might do more to change how people engage with algorithms. For instance, recommender systems may seem more understandable if they are given human characteristics (Waytz, Gray, Epley & Wegner, 2010; Waytz, Heafner, & Epley, 2014), and this might reduce aversion to recommender systems. Or aversion could be reduced if algorithms pause, as if "thinking", before making a recommendation (Buell & Norton, 2011). And allowing people to exert some control over an algorithm's judgment could also enable better sensemaking (Dietvorst, Simmons & Massey, 2016). A more thorough account of the factors that increase subjective understanding could ultimately foster greater trust in algorithmic decisions and recommendations.

It is clear that people judge a recommender system not just by what it recommends, but *how* it recommends. Our work suggests that algorithms can be highly accurate even when they are not modeled after human intelligence. But accuracy alone cannot reduce aversion to algorithms—they need to be understood as well.

## REFERENCES


Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. I*EEE transactions on knowledge and data engineering, 17(6),* 734-749.

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet Recommendation Systems. *Journal of Marketing Research, 37(3),* 363-375.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44(3),* 211-233.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter, 9(2),* 75-79.

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology, 24(4),* 586-607.

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research, 45(1),* 77-93.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101(2),* 127-151.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proc 14th Conf on Uncertainty and Artificial Intelligence,* 43-52.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin, 49(6),* 654-656.

Brynjolfsson, E., Hu, Y., & Simester, D. (2011). Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science, 57(8),* 1373-1386.

Buell, R. W., & Norton, M. I. (2011). The labor illusion: How operational transparency increases perceived value. *Management Science, 57(9),* 1564-1579.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5(4),* 559-583.

Davis, H. L., Hoch, S. J., & Ragsdale, E. E. (1986). An anchoring and adjustment model of spousal predictions. *Journal of Consumer Research, 13(1),* 25-37.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34(7),* 571-582.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243(4899),* 1668-1674.

De, P., Hu, Y., & Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science, 56(11),* 1930-1945.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General, 144(1),* 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science, in press*.

Eggleston, C. M., Wilson, T. D., Lee, M., & Gilbert, D. T. (2015). Predicting what we will like: Asking a stranger can be as good as asking a friend. *Organizational Behavior and Human Decision Processes, 128*, 1-10.

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature, 75(7), 4*50-451.

Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., & Wilson, T. D. (2009). The surprising power of neighborly advice. *Science, 323(5921),* 1617-1619.

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4(2),* 133-151.

Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS), 6(4),* 13.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessments, 12(1),* 19.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM T Information Systems, 22(1),* 5-53.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proc ACM-CSW,* 241-250.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality Social Psychology, 53(2),* 221-234.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of Linear Judgment: A Meta-Analysis of Lens Model Studies. *Psychological Bulletin, 134(3),* 404-426.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7(8),* 368-373.

Koren, Y., & Bell, R. (2011). Advances in collaborative filtering. In *Recommender Systems Handbook,* 145-186 (Springer US).

Krishnan, V., Narayanashetty, P. K., Nathan, M., Davies, R. T., & Konstan, J. A. (2008). Who predicts better?: Results from an online study comparing humans and an online recommender system. In *Proc ACM Conference on Recommender Systems,* 211-218.

Larrick, R.P., & J.B. Soll. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52(1),* 111-127.

Lerouge, D., & Warlop, L. (2006). Why it is so hard to predict our partner's product preferences: The effect of target familiarity on prediction accuracy. *Journal of Consumer Research, 33(3),* 393-402.

Logg, J. (2017). When Do People Rely on Algorithms? *Working Paper.*

Mannes, A. E. (2009). Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Management Science, 55(8),* 1267-1279.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *ACM Human Factors in Computer Systems,* 1097-1101.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence* (University of Minnesota Press).

Müller-Trede, J., Choshen-Hillel, S., Barneron, M., & Yaniv, I. (2017). The wisdom of crowds in matters of taste. *Management Science.*. In press.

Nielsen Company (2015). *Global Trust in Advertising Survey. September 2015.* www.nielsen.com

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84(3),* 231-265.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56-58.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26(5),* 521-562.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proc ACM-WWW,* 285-295.

Sharma, A., & Cosley, D. (2015). Studying and Modeling the Connection between People's Preferences and Content Sharing. In *Proc ACM CSW,* 1246-1257.

Sinha, R. R., & Swearingen, K. (2001). Comparing Recommendations Made by Online Systems and Friends. In *Proc DELOS-NSF: Personalisation and recommender systems in digital libraries.*.

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook* (pp. 479-510). Springer US.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14(8),* 383-388.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52,* 113-117.

Yaniv, I. (2004). The benefit of additional opinions. *Current directions in psychological science, 13(2),* 75-78.

## FIGURE LEGENDS

**Figure 1.** Accuracy results from Studies 1 & 2 comparing human recommendations and machine recommendations (error bars represent standard error of the mean).

**Figure 2.** Recipients' evaluations of the recommenders' ability from Studies 3 & 5, based on perceived recommendation source and recommender description, respectively. (error bars represent standard error of the mean).

**Figure 3.** Recipients' rated understanding of the recommendations from Studies 4 & 5, based on perceived recommendation source and recommender description, respectively. (error bars represent standard error of the mean).

**FIGURE 1**

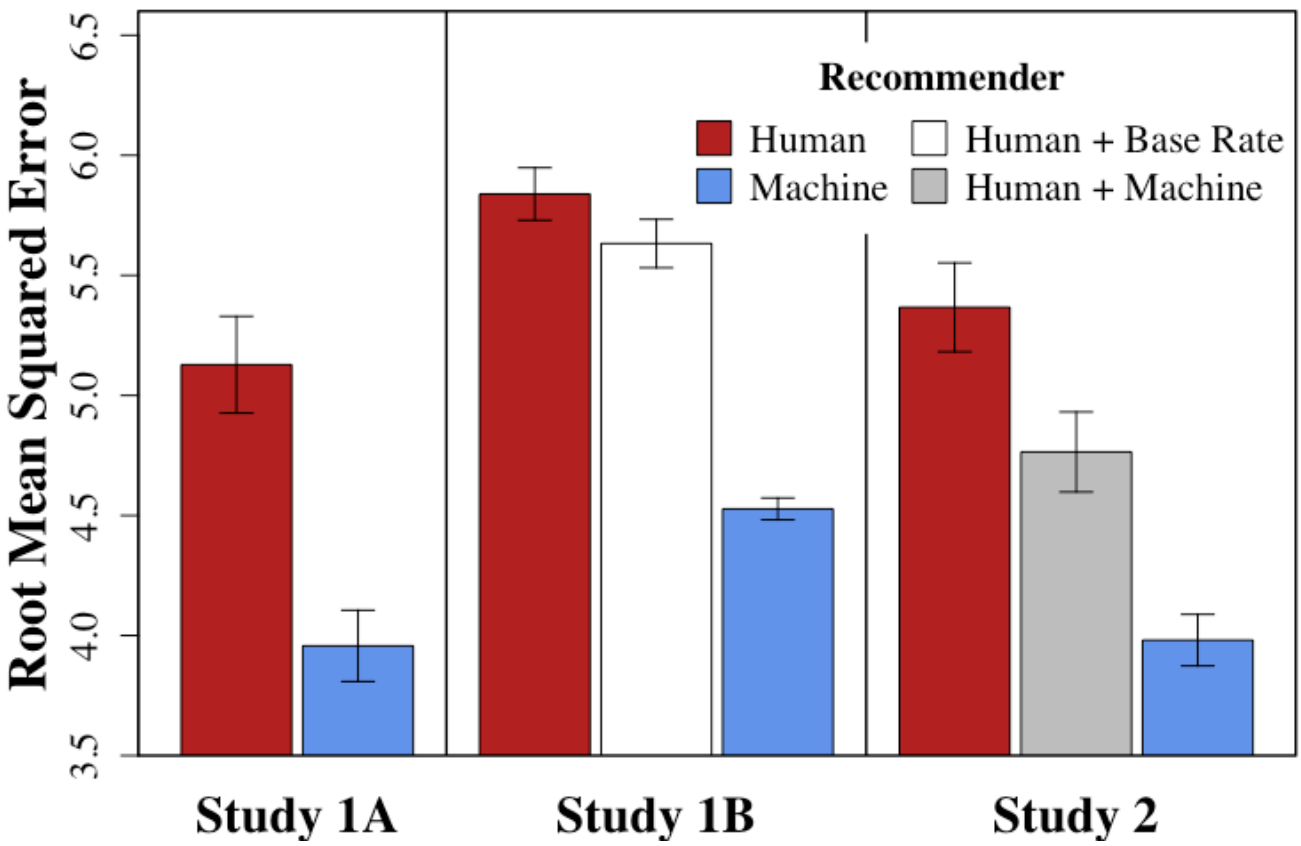

Root Mean Squared Error
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Recommender
Human
Machine
Human + Base Rate
Human + Machine
Study 1A
Study 1B
Study 2

**FIGURE 2**

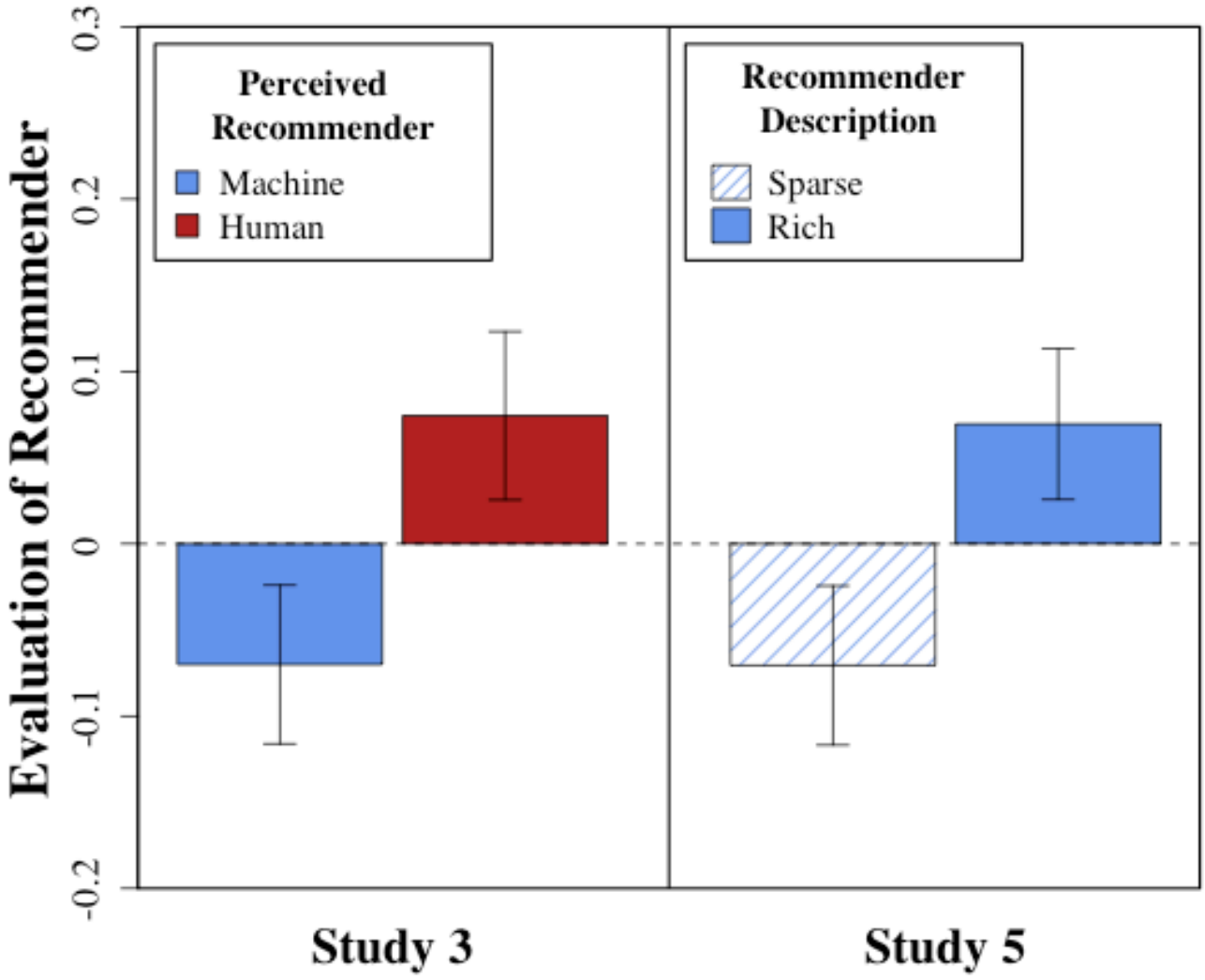
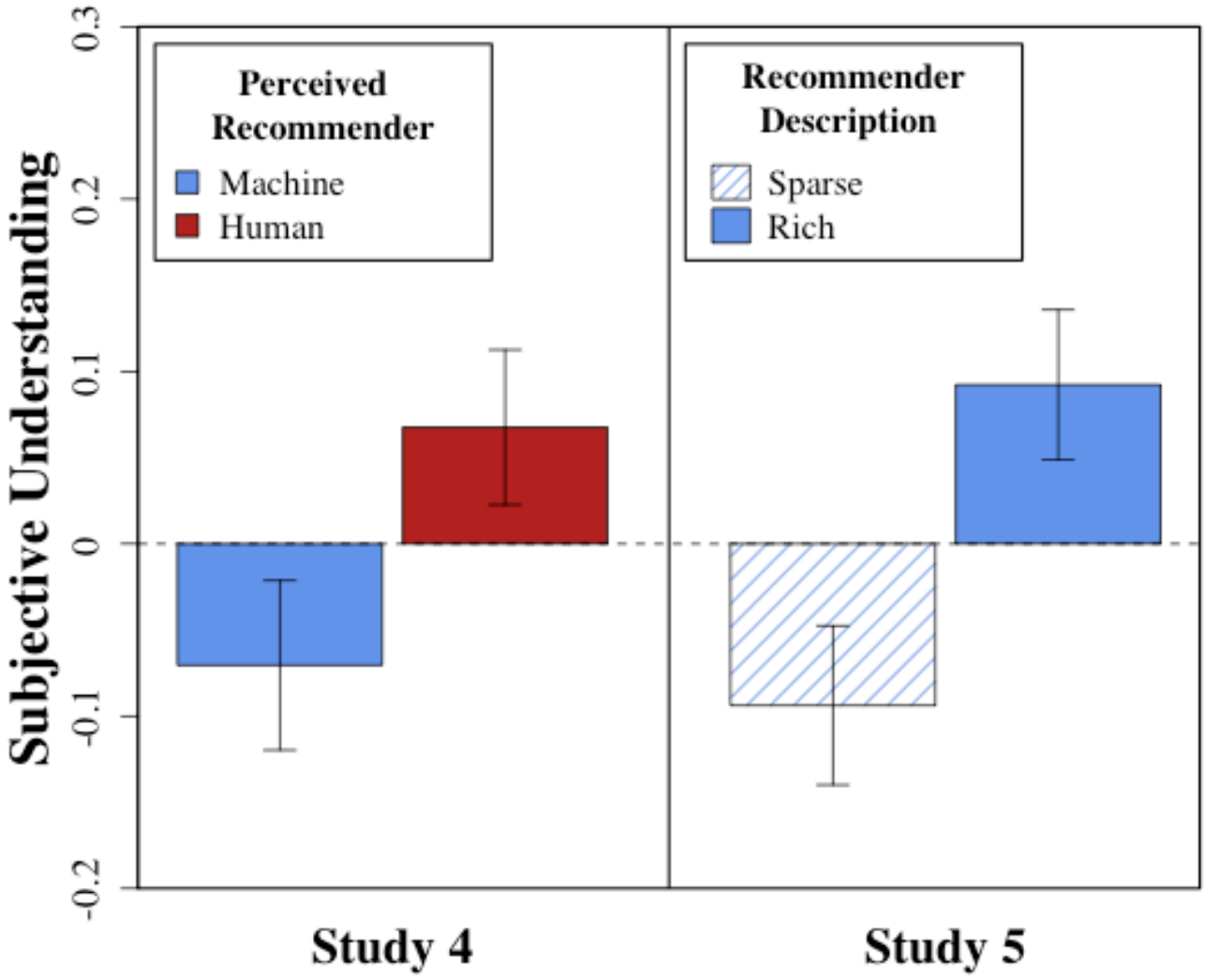

Evaluation of Recommender
0.3
0.2
0.1
0
-0.1
-0.2
Perceived Recommender
Machine
Human
Recommender Description
Sparse
Rich
Study 3
Study 5

**FIGURE 3**


Subjective Understanding
0.3
0.2
0.1
0
-0.1
-0.2
Perceived Recommender
Machine
Human
Recommender Description
Sparse
Rich
Study 4
Study 5

# Supplemental Material

The main text of this paper did not provide full details about two aspects of the experimental methods, for brevity and clarity. However, in the spirit of transparency and full disclosure, we report these details in full here.

First, the main text focused exclusively on the final sample for analyses, after all exclusion criteria have been applied. We used consistent *a priori* exclusion criteria for every study, and below we report our intended and actual sample sizes, before and after exclusions, for every study.

Second, some of the studies in this paper included secondary dependent measures that were not described in the main text. None of these measures have any substantive effect on the interpretation of our primary results, and below we report all of our dependent measures from every study.

## *Sample Size Determination*

Participants recruited from Mechanical Turk were not allowed to complete the study if they had participated in any earlier study we had run (indicated by their worker ID). Additionally, they were not counted in our recruitment total if they failed to complete the study. This includes people who failed the initial attention check (see Appendix B), since they were not allowed to complete the rest of the study afterwards. Among those who were recruited and completed the study, we checked to make sure that they had not given the same rating to every joke (which indicated either lack of effort or total anhedonia), and had passed the manipulation check (for Studies 3-4, see Appendix C).

When participants were recruited at the Museum of Science and Industry, we used no attention checks or manipulation checks. Instead, we excluded participants if they did not complete the study, or if the research assistant (blind to condition) noted that they were severely confused or broke from the protocol through the study. These notes were tabulated in full before we conducted any of our main

analyses.

**Study 1.** We intended to recruit 150 participants (75 pairs) from the Museum of Science and Industry. We actually recruited 150 participants, and 28 participants (14 pairs) were dropped due to incomplete responses or not following instructions, leaving 122 participants (61 pairs).

**Study 1B.** We intended to recruit 500 participants from Mechanical Turk, and 581 were able to complete the study. 34 participants failed the manipulation check and 3 participants gave the same rating to every joke, leaving 544 participants for the analyses.

**Study 2.** We intended to recruit 250 participants (125 pairs) from the Museum of Science and Industry. Due to scheduling conflicts, we actually recruited 232 participants (116 pairs). Twenty-two participants (11 pairs) were dropped due to incomplete responses or not following directions, leaving 210 participants (105 pairs).

**Study 3.** We intended to recruit 1000 participants from Mechanical Turk, and 996 were able to complete the study. 104 participants failed the manipulation check and 6 participants gave the same rating to every joke, leaving 886 participants for the analyses.

**Study 4.** We intended to recruit 1000 participants from Mechanical Turk and 1010 were able to complete the study. 107 participants failed the manipulation check and 4 participants gave the same rating to every joke, leaving 899 participants for the analyses.

**Study 5.** We intended to recruit 1000 participants from Mechanical Turk and 1014 were able to complete the study. 24 participants failed the manipulation check and 4 participants gave the same rating to every joke, leaving 986 participants for the analyses.

***Dependent Measures***

**Study 1.** The primary measures in this study were the ratings participants gave to the jokes, and the predictions they make about their partner's ratings. Participants started by giving their own rating to all twelve jokes, by answering the question "How funny do you think this joke is?" on a continuous scale ranging from -10 ("less funny") to +10 ("more funny"). When it came time to predict their partner's ratings, they answered the question "How funny did your partner think this joke was?" on the exact same -10 to +10 scale.

At the end of the study, several exploratory measures were included to assess participants' knowledge and confidence related to the task, and the order of these measures was randomized. One question asked "Think, in general, about how well you know your partner's taste in jokes. On the scale below, tell us how much you think you know!" and participants responded on a seven-point scale, anchored with "not at all well" and "extremely well". Another question asked "Think, specifically, about the 8 predictions you made for your partner. On the scale below, tell us how many predictions were correct (correct is defined as +/- 2 of their actual rating)", and participants responded with a number from 0 to 8.

As a check that participants knew one another, they were asked "How long have you known your partner, in years?" and gave the following response options: 0-1; 1-2; 2-5; 5-10; 10-20; or 20+. We also asked them "How do you and your partner know each other?" and with the following response options: Spouse; Fiancee; Significant Other; Immediate Family; Extended Family; Work Colleagues; Friends; or Other.

Participants also answered two questions in which they compared their accuracy (and their partners' accuracy) to a recommender system's accuracy. The full text of those questions was:

"We'd like you to imagine a computer algorithm that has a database of people who rated all the jokes you just saw, and can use that database to predict which jokes someone would like (similar to

recommendations at Amazon.com, or Netflix). Now imagine we also told that algorithm what ratings [you/your partner] gave for the four sample jokes, and it tried to predict what ratings [you/your partner] gave on the other eight. How accurate would the computer's predictions be, compared to [your partner's/your] predictions?"

The response was a binary forced choice, between "[I/ my partner] would beat the computer" and "The computer would beat [me/my partner]". The results of this question showed that roughly half of participants chose the recommender system. However, we were concerned that this experiment was conducted in a museum that was devoted to scientific and technological marvels, which may have created demand characteristics that led participants to say they trusted machines, and did not think this data had much insight into people's behavior. Furthermore, people did not have any experience with the recommender system, so they could not evaluate the accuracy, except by perhaps using the prestige brand names ("Amazon", "Netflix") as a cue. Studies 2-5 show that when trust in an algorithm is measured more subtly, after participants have some experience with the machine, people do not use the recommender system very much.

**Study 1B.** The primary measures in this study were the predictions participants made about their targets' ratings. Like Study 1, participants made those predictions by answering the question "How funny did [person] think this joke was?" on a continuous scale ranging from -10 ("less funny") to +10 ("more funny"). For clarity, each of the five targets was differentiated with a letter - "person A", "person B", and so on. Afterwards, participants also gave their own rating to every joke, on the same -10 to +10 scale.

Exploratory measures were included to assess participants' beliefs about a recommender system's accuracy, compared to their own accuracy and to another participants' accuracy. The full text

of those questions was:

"We'd like you to imagine a computer algorithm that has a database of people who rated all the jokes you just saw, and can use that database to predict which jokes someone would like (similar to recommendations at Amazon.com, or Netflix). Imagine we had that algorithm [predict what jokes you would like/ make the same predictions you just did] - that is, it would see [your] ratings of four sample jokes [by each person], and use them to predict [your/their] ratings for other test jokes. [Also, imagine another person in this study tried to predict what jokes you would like - that is, the other person would see your ratings of four sample jokes, and use them to predict your ratings for other test jokes.] How accurate would the computer's predictions be, compared to [your/another person's] predictions?"

The response was a binary forced choice, between "[I/ the person] would beat the computer" and "The computer would beat [me/ the person]". The results of this question showed that roughly half of participants chose the recommender system. However, we were again concerned about the responses, for the same reasons as in Study 1, so we did not put much faith in these responses.

**Study 2.** The primary measures in this study were the ratings participants gave to the jokes, and the predictions they make about their partner's ratings. Participants started by giving their own rating to all twelve jokes, by answering the question "How funny do you think this joke is?" on a continuous scale ranging from -10 ("less funny") to +10 ("more funny"). When it came time to predict their partner's ratings, they answered the question, "How funny did your partner think this joke was?" on the exact same -10 to +10 scale.

All participants, in all conditions, answered the following question about their subjective knowledge: "Think, in general, about how well you know your partner's taste in jokes. On the scale

below, tell us how much you think you know!" and they responded by completing the prompt "I know their sense of humor…" on a 1 to 7 scale, with the anchors "…not at all well" and "…extremely well". As a check that participants in the "known" condition knew one another, they were asked "How long have you known your partner, in years?" with the following response options: 0-1; 1-2; 2-5; 5-10; 10-20; or 20+. We also asked them "How do you and your partner know each other?" with the following response options: Spouse; Fiancee; Significant Other; Immediate Family; Extended Family; Work Colleagues; Friends; or Other.

In the conditions where participants saw the machine predictions, they were also asked two exploratory questions about their confidence in the recommender. The first question was "Think, specifically, about the 8 predictions that you and the algorithm both made for your partner. On all 8 of these predictions, it has to be the case that either your prediction, or the algorithm's prediction, was closer to your partner's true rating (no ties). How many of your predictions do you think were more accurate than the algorithm's predictions?" and participants responded using a number from 0 to 8. They were also asked "Think, in general, about how well the algorithm knew your partner's taste in jokes. On the scale below, tell us how much you think the algorithm knows" and they responded by completing the prompt "I know their sense of humor…" on a 1 to 7 scale, with the anchors "…not at all well" and "…extremely well".

**Study 3.** Human recommenders were collected first, in a separate study. After the attention check, they were told they would make joke recommendations for three different targets. For each target, they read their ratings on three sample jokes, and were then presented with a list of ten jokes. Participants were then told, "From the list of ten jokes below, pick the three that you think they would enjoy the most!" After they chose those jokes, they were given a text box to explain their choice, along

with the following prompt: “Use the box below to tell this person why you thought they would like the jokes you picked! Remember, they will see this text later so make sure you give a thoughtful answer.”

In the main study, the primary measures were the ratings participants gave to the ten jokes and the subjective preference ratings they gave to the recommender. Participants rated each joke on a continuous sliding scale from -10 (“not funny at all”) to +10 (“extremely funny”). The three subjective preference measures were collected on seven-point scales, presented in a random order, with endpoints labelled “not at all” and “extremely”.

“How good do you think the recommender was at choosing jokes you would enjoy?”

“How well do you think the recommender knew your sense of humor?”

“How much would you want to read more jokes that the recommender chose for you?”

**Study 4.** The human and machine recommendations were reused from Study 3. In the main study, the primary measures were the ratings participants gave to the ten jokes and the subjective ratings they gave to the recommender. Participants rated each joke on a continuous sliding scale from -10 (“not funny at all”) to +10 (“extremely funny”). The first two subjective ratings were made on a seven-point scale with the endpoints labelled “strongly disagree” and "strongly agree”, and participants reported their agreement with the following two statements:

“I could understand why the recommender thought I would like those jokes”

“It is hard for me to explain how the recommender chose those jokes”

Finally, participants were asked to make a binary choice between two potential recommenders - “an algorithm” and “another person” - if they were to receive more joke recommendations later on. These options were ordered based on the participant’s condition, so that the recommender they had for the first part of the study was always presented first.

**Study 5.** All recommendations were generated by the same algorithm as in Study 3 & 4, though the sample joke ratings were not rounded. In the main study, the primary measures were the subjective ratings they gave to the recommender.

The explainability measures were collected on seven-point scales, in a random order, with the endpoints labelled “strongly disagree” and "strongly agree”, and participants reported their agreement with the following two statements:

“I could understand why the recommender thought I would like those jokes”

“It is hard for me to explain how the recommender chose those jokes”

The three preference measures were collected on seven-point scales, presented in a random order, with endpoints labelled “not at all” and “extremely”.

“How good do you think the recommender was at choosing jokes you would enjoy?”

“How well do you think the recommender knew your sense of humor?”

“How much would you want to read more jokes that the recommender chose for you?”

## Appendix A: Attention Check

This was used at the beginning of Studies 1B, 3, 4 & 5, and the pilot study. Participants who did not pass the attention check were not allowed to enter the main part of the study, and were not counted in our recruitment totals.

---

First, tell us about yourself!

To help us understand how people think about different activities, please answer this question correctly. Specifically, we are interested in whether you actually take the time to read the directions; if not, the results would not be very useful. To show that you have read the instructions, please ignore the items below about activities and instead type 'I will pay attention' in the space next to 'Other'. Thank you.

[ ] Watching Athletics
[ ] Attending Cultural Events
[ ] Participating in Athletics
[ ] Reading Outside of Work or School
[ ] Watching Movies
[ ] Travel
[ ] Religious Activities
[ ] Needlework
[ ] Cooking
[ ] Gardening
[ ] Computer Games
[ ] Hiking
[ ] Board or Card Games
[ ] Other: _____

## Appendix B: Manipulation Check

This was used at the end of Studies 3, 4 & 5 to confirm that participants had processed the information they were given about the source of the recommendations they had received.

---

Answer a quick question about the experiment you just took part in, to make sure you were paying attention.

How were the final three jokes you saw chosen?

[ ] Another person in this experiment
[ ] Someone from a different experiment
[ ] A recommendation algorithm
[ ] A random choice

**Appendix C : Screenshot from Study 1**

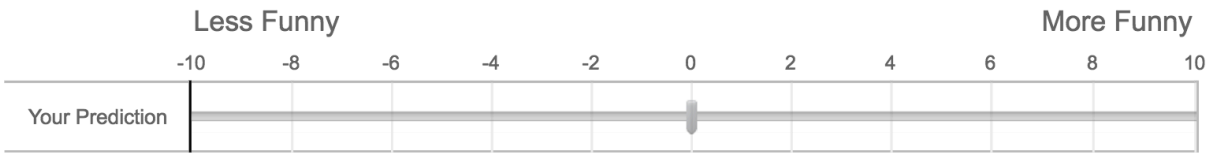
On the left are sample jokes, with the target's ratings. On the right is the test joke. Below, participants use the slider to predict how much their partner will like the test joke. We have removed the text of the jokes here, though **a list of all jokes used in all studies are provided online.**

*For reference, here are some jokes your partner rated...*

| Sample Joke 1 | Sample Joke 2 |
|---|---|
| [ Text of sample joke 1] Your partner rated this joke -8.6 | [ Text of sample joke 2] Your partner rated this joke -7.9 |
| **Sample Joke 3** | **Sample Joke 4** |
| [ Text of sample joke 3] Your partner rated this joke 4.3 | [ Text of sample joke 4] Your partner rated this joke -3.6 |

*How funny did your partner think this joke was?*

[ Text of test joke ]



## Appendix D: Explicit Choice Prediction Study

### Methods

Two hundred and one participants from Amazon.com's Mechanical Turk (MTurk) platform completed our study. Four failed the attention check, leaving 197 participants for the analyses. Participants served only as recommenders, not targets.

The targets were not new participants, but rather they were people who had previously rated jokes in an online database. We used data from a subset of 5520 people who had all rated the same 30 jokes (Goldberg et al., 2001). Most of these people ($N$ = 4520) were used as the training set for the collaborative filtering algorithm (i.e., to estimate the correlations between joke ratings for all 30 jokes). However, we randomly selected 1000 people to form a "holdout set." The holdout set was our pool of potential targets.

Each participant in our study (who served as recommenders) made recommendations for five targets from the holdout set. For each target, participants first saw four sample jokes and the target's ratings of those jokes. To make recommendations, participants then saw two new test jokes. Participants picked which of these two jokes they thought the target rated higher. They did this for all five targets, and no jokes were repeated, so all 30 jokes were used exactly once for every participant. However, the order of jokes, including which jokes were included in the sample set, was determined randomly for every participant. Accuracy was incentivized by giving a $20 bonus to the most accurate participant. At the end of the study, participants personally rated each joke.

For each pair of test jokes, the recommender system predicted the target's rating for each joke. We then inferred that the system "recommended" the joke with the higher predicted rating (e.g., if the system predicted Joke A would receive a "6.2" and Joke B would receive a "4.3" then that was coded as the system recommending Joke A).

**Results**

Accuracy was scored as the percentage of times a recommender correctly guessed which test joke the target rated higher (random guessing would score 50%). Human recommenders ($M$ = 56.6%, *95% CI* = [53.6%, 59.7%]) were more accurate than random guessing, and were also more accurate than if they had just picked their own favorite joke from every pair ($M$ = 50.4%, *95% CI* = [47.2%, 53.5%]), $t(196) = 4.02$, $P < .001$. However, the machine ($M$ = 62.9%, *95% CI* = [59.8%, 66.1%]) again outperformed human recommenders, $t(196) = 3.15$, $P = .002$.

## Appendix E: Recommender System Explanations

The two conditions in Study 5 differed only in the amount of explanation that participants received. This difference was operationalized on three pages in the survey: the introduction page; the page on which participants waited for the recommended jokes; and the page on which participants were shown their recommended jokes.

Below, we show exactly how those pages differed between conditions. On each page, the entire text of the sparse condition was shown in both conditions, but in the rich condition, an additional explanation was added. Here, we add italics to show which part of the text was only shown in the rich condition - however, these italics were not part of the actual stimuli.

---

### Introduction Page

In this study you will receive recommendations from a recommendation algorithm. First, you are going to read three jokes and rate how funny you think they are. Then we are going to feed your ratings into a computer algorithm, which will recommend three jokes that you might also like.

*Here's how the algorithm works. The algorithm uses a database of other people's ratings of different jokes, including three sample jokes you will rate.*

*Think of the algorithm as a tool that can poll thousands of people and ask them how much they like different jokes. This way, the algorithm can learn which jokes are the most popular overall, and which jokes appeal to people with a certain sense of humor.*

*Using the database ratings, the algorithm will search for new jokes that are similar to the ones you liked, and dissimilar to the ones you did not like. The algorithm will then recommend some new jokes you might like.*

### Waiting Page

Your ratings for those three jokes were sent to the algorithm, which will search a database of jokes rated by thousands of people, that includes the sample jokes you just rated.

*Right now, this algorithm is using your ratings to guess which new jokes you might like.*

*Think of the algorithm as a tool that is polling thousands of people and asking them how much they like different jokes. The algorithm is learning which jokes are the most popular overall, and which jokes are appealing to people with your sense of humor.*

*The algorithm will choose some new jokes to show you by searching for new jokes that are similar to the ones you liked, and dissimilar to the ones you did not like.*

*Here is the input you gave to the algorithm:*

| ***JOKE*** | ***RATING*** |
|---|---|
| *[sample joke 1]* | *[rating 1]* |
| *[sample joke 2]* | *[rating 2]* |
| *[sample joke 3]* | *[rating 3]* |

**Recommendation Show Page**

These are the jokes that the algorithm chose for you.

We'd like you to read each one, and rate how much you like each one,
on the same scale as before - from -10 (not funny at all) to 10 (extremely funny).

| **JOKE** | **RATING** |
|---|---|
| [recommended joke 1] | [slider] |
| [recommended joke 2] | [slider] |
| [recommended joke 3] | [slider] |

***Explanation:*** *The algorithm selected these jokes because most people who rated the first three jokes like you did also liked these jokes. That is, these jokes are popular among people who give ratings that are similar to the ratings you gave:*

| ***JOKE*** | ***RATING*** |
|---|---|
| *[sample joke 1]* | *[rating 1]* |
| *[sample joke 2]* | *[rating 2]* |
| *[sample joke 3]* | *[rating 3]* |

After you've read all three jokes, press “Continue”.