# Making Sense of Recommendations

Jon Kleinberg
Cornell University

Sendhil Mullainathan
Harvard University

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

Mike Yeomans
Harvard University

Department of Economics
Littauer Center
1805 Cambridge Street
Cambridge, MA 02138
yeomans@fas.harvard.edu

**Abstract.** Algorithms are increasingly being used to make recommendations about matters of taste, expanding their scope into domains that are primarily subjective. This raises two important questions. How accurately can algorithms predict subjective preferences, compared to human recommenders? And how much do people trust them? Recommender systems face several disadvantages: They have no preferences of their own and they do not model their recommendations after the way people make recommendations. In a series of experiments, however, we find that recommender systems outperform human recommenders, even in a domain where people have a lot of experience and well-developed tastes: Predicting what people will find funny. Moreover, these recommender systems outperform friends, family members, and significant others. But people do not trust these recommender systems. They do not use them to make recommendations for others, and they prefer to receive recommendations from other people instead. We find that this lack of trust partly stems from the fact that machine recommendations seem harder to understand than human recommendations. But, simple explanations of recommender systems can alleviate this distrust.

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

**Significance Statement.** As data science broadens in scope, algorithms are increasingly being used to predict matters of preference and make recommendations about what people will like. Subjective predictions pose important computational challenges: Algorithms are often blind to content, and they cannot simulate human tastes. As a result, these systems are often thought of as cheap substitutes for human recommenders and people do not trust them. Here, we find that these systems outperform humans, even when humans are afforded many advantages over recommender systems. And simply explaining how the recommender systems operate can foster trust in them. This work demonstrates how a better understanding of people's interaction with algorithms can support better decision-making at scale.

**Disclosure.** 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 Supporting Information at https://osf.io/8nbr2/

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 [1]. Most notably, recommender systems now predict which movies and books people will enjoy, which products they should buy, and which restaurants they should visit [2].

Although recommender systems are now widespread in practice, people are reluctant to trust them. In a pilot survey, we asked 100 people to use an algorithm that would learn which jokes they found funny (we used this domain because humor is highly subjective) [3]. The algorithm then recommended additional jokes from a database to participants. After reading the recommended jokes, people then indicated whether they thought the algorithm or a person would perform better at recommending jokes. Participants overwhelmingly thought that other people would do a better job of recommending jokes from the database—69% of participants thought another person would be more accurate than the algorithm, and 74% preferred to receive recommendations from people instead of the algorithm.

Their skepticism is understandable. After all, even though recommender systems are becoming ubiquitous, we are still accustomed to receiving recommendations from other people. 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 we trust this process—83% of people trust recommendations from friends and family; 66% trust online opinions of strangers [4]. We trust it for good reason. The opinions of other people are surprisingly powerful predictors of what we will like in the future, even compared to our own predictions [5,6].

Furthermore, human recommenders possess a wealth of knowledge that recommender systems do not have. Humans know the content of what they are recommending, can draw on their own experiences, are often familiar with the recipient, and may know the context in which recommendations will be experienced. By contrast, recommender systems operate in the dark. They have limited

information about the recipient, and no direct knowledge of what they are recommending. They cannot read a book, watch a movie, or taste the food at a restaurant. They only know *what* we like, not *why* we like it.

Simply put, recommender systems cannot mimic the human mind. This contrasts with previous statistical ("actuarial") models that have been shown to improve on human (i.e., "clinical") judgment by mimicking and producing a more consistent version of the same process [1,7,8,9]. Those models use the same kinds of information that humans do and are built on the assumption that, 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" [7]. Recommender systems often lack this special expertise, so it is no surprise that they might be viewed as a cheap substitute for human judgment.

This raises two natural questions. First, how accurate are recommender systems, compared to human recommenders? Second, what influences people's trust in recommender systems? Here we provide a novel and thorough test of these questions in a subjective domain where humans should have a strong advantage: predicting which jokes people will like. There are several reasons to think that humor is good proxy for other matters of taste. Humor is a universal element of social interaction [10]. And predicting what a person will find funny seems to require (a) deep, almost intimate, knowledge of the person and (b) a theory of why certain jokes would match someone's sense of humor. In domains like this, humans have an enormous a priori advantage because they can use knowledge and theory that is inaccessible to a recommender system. Moreover, humans regularly get clear, immediate feedback on what people find funny, which makes it easier to develop expertise [11].

Attempts to compare the accuracy of human and machine recommendations have thus far been inconclusive [12,13,14]. This is because no study has actually measured participants' direct experience of the recommended items. In one study, participants were merely asked to *predict* how much they would like a recommended item, but such forecasts are notoriously unreliable [12,15]. And other work

did not ask participants to experience new items. Instead, the recommenders predicted which of participants' *past* experiences they had enjoyed most at the time. Essentially, the recommenders were "recommending" experiences that participants had already chosen for themselves, which tells us little about how well recommenders fare at suggesting novel experiences [16]. In contrast, our work compares human and machine recommenders in a controlled setting, where all participants could directly experience the same set of new items. And we conducted these experiments with large samples, with varying degrees of familiarity between the people making and receiving recommendations.

We have three main findings. First, recommender systems consistently outperform human recommenders, whether these people are strangers, friends, family members, or significant others. Second, despite this robust advantage, people trust recommender systems less than they trust humans. For instance, when making recommendations for others, people do not effectively use input from recommender systems—even when the algorithmic recommendations are outperforming their own. And people prefer recommendations which they believe come from a human (instead of a recommender system). In a sense, even when people liked the system's *recommendations*, they did not like the system as a *recommender*. Finally, we explore why people distrust recommender systems. We find that people think that recommendations from a system are harder to understand than recommendations from a person [17,18]. But, when people are given explanations about how the recommender system operates, their trust in the system increases.

Although there is a vast array of recommender systems available today, we focus on the most typical class known as "collaborative filtering." These systems exclusively use data on how much people enjoyed past experiences [19,20]. While complex hybrid systems have also been developed - such as using human annotators, or parsing the text of the jokes, or scraping social media profiles - collaborative filtering is a core component of every recommender system, and its performance can be more clearly contrasted against the power of human recommendations. Furthermore, by forgoing more

complicated algorithms we provide a conservative test of the accuracy of recommender systems.

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, (2) they did not complete the study, (3) they did not follow instructions, (4) they failed a manipulation check, 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 Supporting Information.

## 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 - family, friends, partners, and so on.

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 saw 12 jokes 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 switched computer terminals, where they saw their partner’s ratings for four of the jokes (the “sample set”), randomly selected from the full set. They then predicted their partner’s ratings for the remaining eight jokes (the “test set”).

We then compared the accuracy of each participant’s predictions to the accuracy of a recommender system’s predictions. The system predicted the target’s ratings of the eight test jokes based on their sample joke ratings. We also estimated the accuracy of “surrogation” information [5, 21], by testing how accurate human recommenders would have been if they simply predicted that their partner would like a joke as much as they did.

The recommender system used an item-based collaborative filtering algorithm, which was trained on a database of ratings from 454 participants pooled from this study and a similar previous study in the same setting with the same jokes. This algorithm essentially models the relationship between the sample jokes and test jokes using a simple least-squares regression [22]. While there is a long literature on computational approaches to recommendations, this method is common and represents a middle-ground in terms of complexity. [2]. We used a leave-one-out cross-validation procedure to make each prediction. That is, for every target, the model was trained on ratings from the

453 other participants in the database.

**Results**

To measure accuracy, we constructed all 28 possible pairwise comparisons from the set of eight test jokes. If a recommender gave a higher rating to the item in the pair that the target enjoyed more, 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 (see Figure 1). This ensures that participants were not penalized for using the scale less effectively than the recommender system. Human recommenders ($M$ = 56.8%, $SE$ = 1.4%) marginally outperformed surrogation ($M$ = 54.6%, $SE$ = 1.2%; paired t-test: $t(121) = 1.7$, $P = .084$), suggesting they had some insight into their targets' preferences. However, they were significantly less accurate than the machine recommender system ($M$ = 61.1%, $SE$ = 1.2%; paired t-test: $t(121) = 2.6$, $P = .009$).

To our knowledge, this is the first test comparing recommender systems to human recommenders in which participants all saw the same set of novel items. In this design, our recommender system even outperformed people who know each other well. But perhaps participants' judgments were *clouded* by their intimate knowledge of their partners [23,24]. Perhaps the objectivity of strangers might make for better recommendations. And it is also possible that humans are more adept at comparing jokes than predicting each rating in isolation. The next study addresses these issues.

## STUDY 1B

### 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 selected from an online database of people who rated jokes [3]. We used data from a subset of 5,520 people who had all rated the same 30 jokes. One thousand people were randomly selected from the dataset as a "holdout set" to form a pool of targets for the recommenders, and the data from the remaining individuals ($N$ = 4,520) were used as the training set for the collaborative filtering algorithm.

Participants made recommendations for five randomly selected targets. For each target, participants first saw four sample jokes and the target's ratings of those jokes. Participants then picked which of two new test jokes they thought the target rated higher (example stimuli in Appendix A). They did this for all five targets, and no jokes were repeated. 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 generated a separate prediction of how much the target would like each test joke (which was coded as choosing the joke with the higher predicted rating).

### 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%, $SE$ = 1.5%) were again able to exceed the accuracy of surrogation ($M$ = 50.4%, $SE$ = 1.5%; paired t-test: $t(196) = 4.0$, $P < .001$). However, the machine ($M$ = 62.9%, $SE$ = 1.6%) again outperformed human recommenders (paired t-test: $t(196) = 3.1$, $P = .002$).

These recommender systems outperform both strangers and close partners. But do people trust

these systems? On the one hand, the increasing popularity of these systems suggests that, on some level, people do trust them. But it is also possible that people simply see them as cheap substitutes for human judgment. In the next study, we test whether people are willing to have recommender systems complement human judgment, by showing them the machine recommendations before they recommend for other people.

## STUDY 2

### *Methods*

As in Study 1A, we recruited 232 participants in 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 this information.

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 from Study 1A, whom they did not know. Both factors were randomized at the pair level (i.e., people recruited together were always in the same condition).

### *Results*

Accuracy was defined as in Study 1A. Once again, recommender systems ($M$ = 62.8%, $SE$ = 1.0%) outperformed humans ($M$ = 58.1%, $SE$ = 1.0%; $t(209) = 4.1$, $P < .001$), and this was true whether they were assigned to recommend for close others (humans: $M$ = 59.1%, $SE$ = 1.4%; machines:

*M* = 62.7%, *SE* = 1.4%; $t(109) = 2.0$, $P = .043$) or for strangers (humans: *M* = 56.9%, *SE* = 1.4%; machines: *M* = 62.9%, *SE* = 1.3%; $t(109) = 4.0$, $P < .001$).

More importantly, participants with the machine's predictions were not more accurate (*M* = 58.3%, *SE* = 1.4%) than participants without the machine's predictions (*M* = 57.9%, *SE* = 1.4%, $t(208) = 0.2$, $P = .809$, see Figure 1). But participants did not ignore the machine's recommendations completely. Human predictions were in fact more highly correlated with the recommender system when they saw its predictions ($r = .521$), compared to when they did not ($r = .348$; multiple regression interaction term: $\beta = .136$, *SE* = .044; $t(1676) = 3.1$, $P = .002$). This suggests that participants used the algorithm to calibrate their own use of the ratings scale, and perhaps to rein in very extreme ratings. But the machine had little influence on which jokes the participants would have chosen, and did not improve their accuracy.

This study shows that even though the recommender system was more accurate, participants did not trust the recommender system enough to use it. Had their trust been higher, they would have recommended jokes that were better suited to their target. But it is not clear whether participants think recommender systems are poor substitutes for just their own recommendations, or for other people's recommendations, as well. That is, how do people feel about these systems when they are recipients of their recommendations? We turn to this question in the remaining studies.

## STUDY 3

People's impressions of recommenders might depend on two factors: (1) the content of a recommendation (i.e., which jokes are chosen) and (2) the process by which it is recommended. Humans and machines often differ on both dimensions, making it difficult to determine which factor might be more influential in shaping these impressions. In this study, we disentangle these two factors by manipulating the *actual* source of recommendations and the *perceived* source.

### *Methods*

All participants received recommendations from either another person or from our recommender system, based on the ratings they gave to some 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 sample jokes. We rounded participants' sample ratings to the nearest 2.5-point marking on the scale, which meant that each joke 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 (e.g., Sample Joke 1: 2.5, Sample Joke 2: -5.0, Sample Joke 3: 7.5) and then picked the three test jokes that someone with those ratings would like most. Each recommender made three sets of recommendations, and these were all pooled together. This database allowed us to have a human recommendation ready for every participant, no matter how they rated the sample jokes.

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 (Appendix C) and 6 participants gave the same rating to 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 (see Appendix C).

### *Results*

**Accuracy.** We compared recommender accuracy in two ways. First, we conducted within-subjects comparisons of the average rating of the three jokes that the machine and human recommenders picked (or would have picked). The recommender system picked jokes that participants found funnier ($M$ = 3.03, $SE$ = 0.12) than did human recommenders ($M$ = 2.74, $SE$ = 0.12; paired t-test:

$t(885) = 3.0$, $P = .003$).

We also computed accuracy based on pairwise comparisons as follows. For each participant, there were 21 possible pairs of recommender-chosen jokes and recommender-excluded jokes (because recommenders chose three jokes from the ten-joke test set). For each pair, if the recommender-chosen joke had a higher rating than the recommender-excluded joke, then this was counted as a correct choice. Accuracy was computed as the percentage of correct choices across the 21 pairs. By this measure, machine recommenders were again more accurate ($M = 53.4\%$, $SE = 0.7\%$) than humans ($M = 51.6\%$, $SE = 0.7\%$; paired t-test: $t(885) = 2.1$, $P = .035$).

**Preference.** Next, we compared how participants rated the *recommenders.* We conducted multiple regressions to analyze the effect of both the actual and the perceived recommender on how participants judged the recommender. Participants thought their recommender was better at choosing jokes if they *thought* the recommender was a human ($\beta = 0.33$, $SE = 0.11$, $t(883) = 2.9$, $P = .004$). This rating did not depend on whether the jokes were *actually* chosen by a human ($\beta = -0.14$, $SE = 0.11$, $t(883) = 1.3$, $P = .211$). Participants also said the recommender knew their sense of humor better when they thought it was a human ($\beta = 0.21$, $SE = 0.12$, $t(883) = 1.8$, $P = .078$). Again, this judgment did not depend on whether the recommender was actually human ($\beta = -0.06$, $SE = 0.12$, $t(883) = 0.5$, $P = .595$). Desire to see more jokes from the recommender followed a similar pattern (perceived human: $\beta = 0.18$, $SE = 0.12$, $t(883) = 1.5$, $P = .134$; actual human: $\beta = -0.05$, $SE = 0.12$, $t(883) = 0.5$, $P = .649$). We standardized and combined all three responses into a single “preference index” (Cronbach’s $\alpha = 0.95$), which is shown in Figure 2. This analysis confirmed that the perceived recommender had a significant effect on subjective ratings ($\beta = 0.14$, $SE = 0.07$, $t(883) = 2.2$, $P = .031$), while the actual recommender had no effect ($\beta = -0.05$, $SE = 0.07$, $t(883) = 0.8$, $P = .433$).

These results suggest that people prefer human recommenders. This is not due to differences in characteristics of the recommendations, such as serendipity, or diversity [25,26]. In fact, accuracy had a strong positive correlation with the preference index ($r = 0.350$, $t(884) = 11.1$, $P < .001$). Nevertheless

recommender systems were judged more harshly. In a multiple regression 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. Recommender systems are inscrutable, and difficult for people to explain [15,16]. But it may be easier to understand how a human would recommend a joke. We test this hypothesis in the next study.

## STUDY 4

### *Methods*

One thousand ten participants from MTurk completed our study; 107 failed the manipulation check (Appendix C) 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, participants indicated whether they preferred to receive additional recommendations from humans or from the recommender system, which more directly assesses trust in the recommender. 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*

Recommender systems were once again more accurate ($M$ = 54.3%, $SE$ = 0.7%) than human recommenders ($M$ = 51.1%, $SE$ = 0.7%; paired t-test: $t(898) = 3.4$, $P < .001$). The rest of our analyses are collapsed across the *actual* recommender, to focus on the effects of the *perceived* recommender. This way, the actual jokes being recommended are held constant.

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%, $SE$ = 1.9%) than when they started with a human

recommender ($M$ = 68.8%, $SE$ = 2.2%; $\chi^2(1, N = 899) = 12.8$, $P < .001$). Put simply, a majority of participants preferred human recommenders ($M$ = 54.8%, $SE$ = 1.7%, $\chi^2(1, N = 899) = 8.4$, $P = .004$).

The subjective ratings were combined in a single "explainability index" (Cronbach's $\alpha = 0.82$). Participants rated human recommenders as easier to understand ($M$ = 0.07, $SE$ = 0.05) than machine recommenders ($M$ = -0.07, $SE$ = 0.05; t-test: $t(897) = 2.1$, $P = .038$). 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, $\beta = 0.60$, $SE = 0.09$, $z(897) = 7.0$, $P < .001$). And this relationship was attenuated when participants thought their recommender was human ($\beta = 0.43$, $SE = 0.11$, $z(457) = 3.8$, $P < .001$; interaction term: $\beta = -0.39$, $SE = 0.18$, $z(895) = 2.2$, $P = .028$). This suggests that humans recommenders are more trusted because it is easier to understand how humans make recommendations.

The results of this study 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 show that people are less willing to accept recommender systems whose process they cannot understand. Would making machine recommendations easier to understand therefore 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 972 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 explainability 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 D for exact stimuli).

### *Results*

Participants in the *rich* explanation condition rated the recommender system as easier to

understand (*M* = 0.09, *SE* = 0.05) than participants in the *sparse* condition (*M* = -0.09, *SE* = 0.05; independent samples t-test: *t*(984) = 2.9, *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, *SE* = 0.04) than participants in the *sparse* condition (*M* = -0.07, *SE* = 0.05; independent samples t-test: *t*(984) = 2.2, *P* = .028). This effect was significantly mediated by explainability (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, which in turn made the algorithm a more preferable source for recommendations.

## GENERAL DISCUSSION

These experiments evaluate a new approach to a familiar problem: Predicting what people will like. We find that recommender systems outperform human recommenders, despite the fact that these systems are blind to the content of what they are recommending. They cannot watch a movie, read a book, or laugh at a joke. They have no model of *why* or *how* we enjoy the things we do. Instead, they can only draw on a matrix of ratings to make recommendations. And yet the studies above suggest that this limited information alone provides a remarkably powerful guide for making recommendations. They outperform strangers and even people who know each other well.

Despite this fact, people are reluctant to rely on recommender systems to aid their own judgment. And people prefer to receive recommendations from other people rather than these systems. Apparently, a good recommendation does not merely depend on the quality of what is recommended. Rather, it also depends on how easy it is to understand the recommendation process. Because the machine recommendation process is relatively inscrutable, people trust recommender systems less than they trust other people, even though people make more errors.

These findings suggest that people's discomfort with these types of algorithms may run deeper than previously appreciated. In research on algorithms that forecast objective events [7], distrust of algorithms tends to arise only after they make mistakes [27]. But in subjective domains, where the success of a recommender system is in the eye of the recipient, people seem to have a deeper resistance to predictions from a machine [28]. Our research provides evidence that perhaps this greater discomfort stems from simply not understanding how the recommender system operates.

These questions are crucial because recommender systems are already ubiquitous, and the most common focus has been on efforts to engineer more accurate algorithms. This is notably embodied in the "Netflix challenge", a million dollar prize awarded to an algorithm that could improve the accuracy of the company's movie recommendations by 10% [29]. Our results here, though, suggest that this may

not be sufficient for increased consumer satisfaction. In general, people may judge a recommender system not just by what it recommends, but *how* it recommends.

## REFERENCES


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

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

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

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

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

6. 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. *Organ Behav Hum Dec, 128,* 1-10.

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

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

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

10. Martin, R. A. (2010). *The psychology of humor: An integrative approach.* Academic press.

11. Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *Am Psychol, 64(6),* 515-526.

12. 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.*

13. 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 Conf on Rec Sys,* 211-218.

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

15. Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. *Adv in Exp Soc Psych, 35,* 345-411.

16. Marlin, B. M., & Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proc ACM Conf on Rec Sys*, 5-12.

17. Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering

recommendations. In *Proc ACM-CSW,* 241-250.

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

19. Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proc 14th Conf on Uncertain Artif Intel,* 43-52.

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

21. Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *J Pers Soc Psychol, 53(2),* 221-234.

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

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

24. 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. *J Consum Res, 33(3),* 393-402.

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

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

27. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *J Exp Psychol Gen, 144(1),* 114.

28. Logg, J. (2016). When Do People Rely on Algorithms? *Working Paper.*

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

## FIGURE LEGENDS

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

**Figure 2.** Average evaluations of the recommenders' ability from Study 3, based on perceived and actual recommendation source (error bars represent standard error of the mean).

**Figure 3.** People's rated understanding of the recommendations from Study 4 & 5 (error bars represent standard error of the mean).

**FIGURE 1**

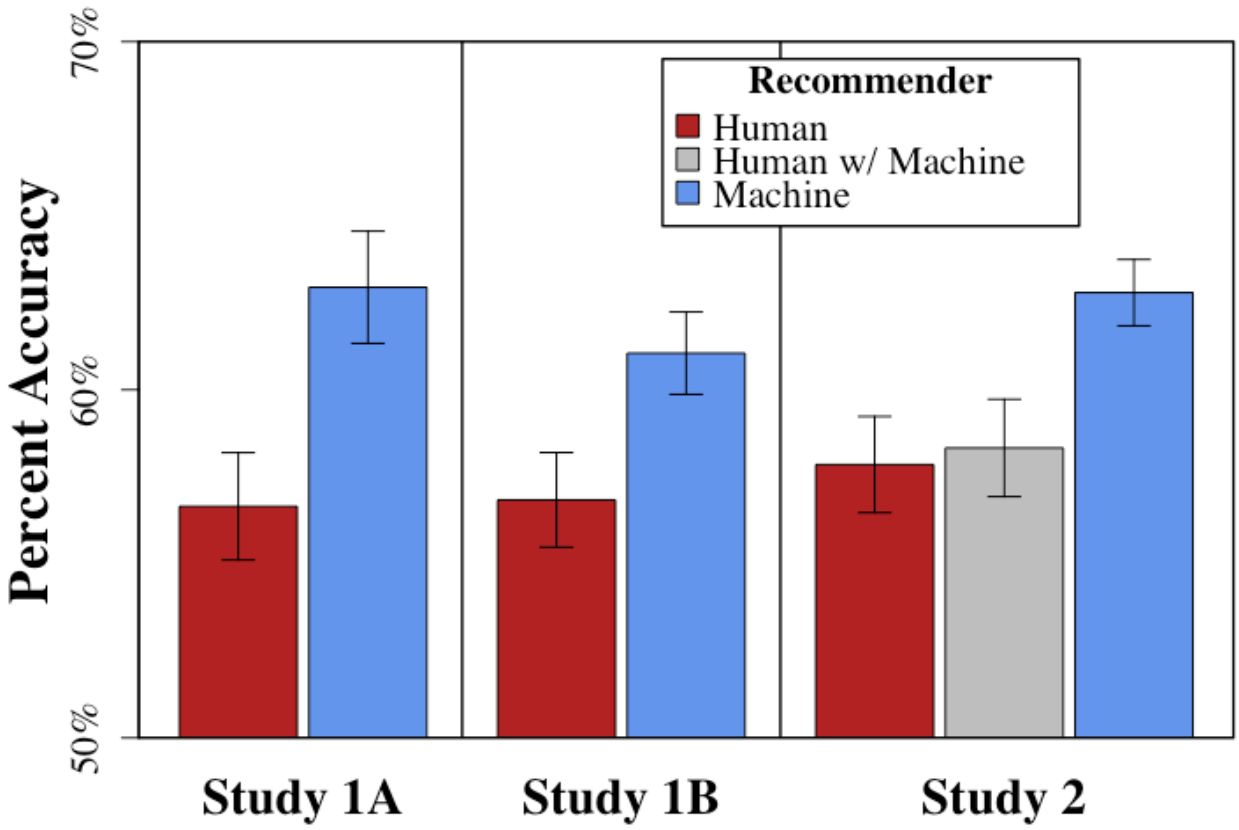

Percent Accuracy
70%
60%
50%
Recommender
Human
Human w/ Machine
Machine
Study 1A
Study 1B
Study 2

**FIGURE 2**

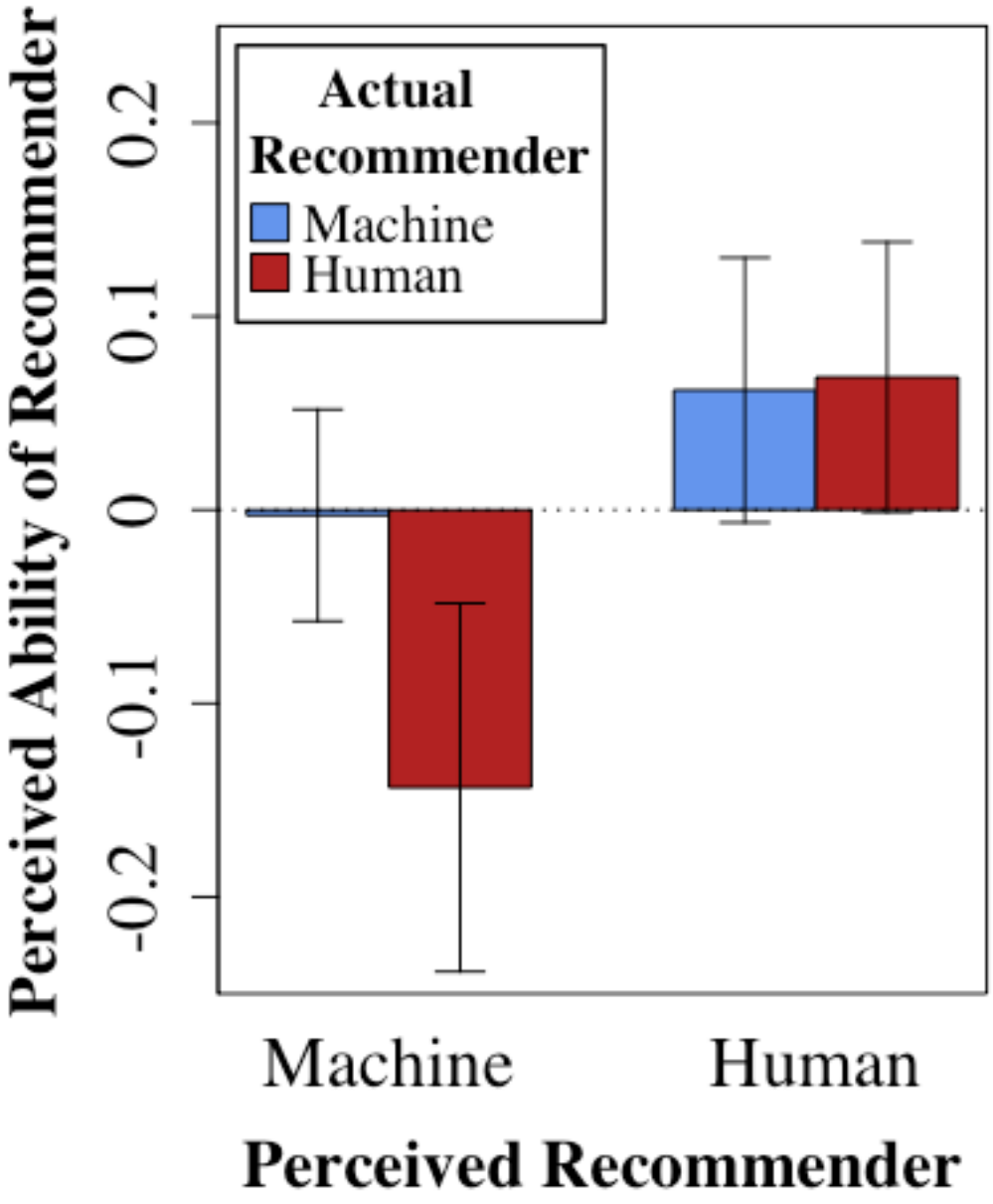
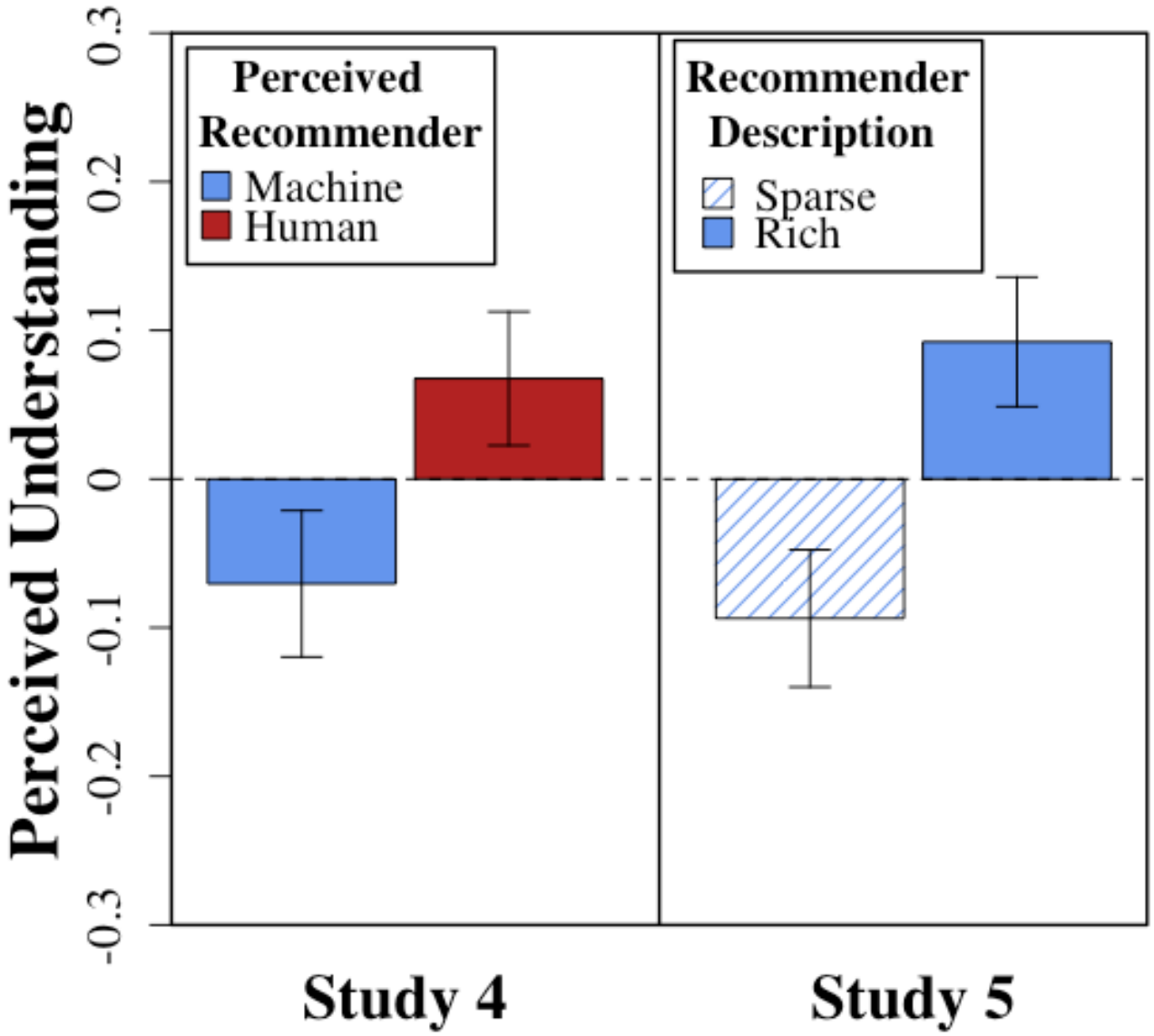

Perceived Ability of Recommender
-0.2
-0.1
0
0.1
0.2
Actual Recommender
Machine
Human
Machine
Human
Perceived Recommender

**FIGURE 3**


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

## Supporting Information

The main text of this paper skirted two important details 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.

**Pilot Study.** We intended to recruit 100 participants from Amazon.com's Mechanical Turk. 102

participants completed the study, and 2 participants failed the attention check, leaving 100 participants for the analyses.

**Study 1A.** 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 200 participants from Amazon.com's Mechanical Turk platform, 201 participants completed the study, and 4 failed the attention check, leaving 197 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 972 participants for the analyses.

***Dependent Measures***

**Pilot Study.** Participants were first sent to a website for a joke recommender system [Goldberg et al], where they were asked to rate eight "sample" jokes and then see five recommended jokes, one at

a time. Participants reported the jokes they saw, and the ratings they gave, to confirm that they had followed instructions, and that they liked the recommended jokes.

Afterwards, participants compared the recommender system to human recommenders, in a 2x2 factorial design. One factor was the kind of human recommender: some participants identified a person "who knew their sense of humor well" for comparison, while other participants compared the algorithm to an unidentified other person. Our primary dependent measure was participants' answers to the following two questions (order counterbalanced):

*Imagine [your friend,* [the person's name]/*another person] also saw your ratings for the first eight jokes and predicted how you would rate the five new jokes, just like the Jester algorithm....*

*...Whose predictions would be closer, on average, to your actual ratings?*

*...Whose recommendations would you prefer to get?*

For every question in all conditions, participants responded to a two-choice forced-alternative between the options "Another Person" or "Jester Algorithm."

**Study 1A.** 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. Our final study shows that, in fact, when trust in an algorithm is measured more subtly in this same context, participants do not use the recommender system very much.

**Study 1B.** The primary measures in this study were the five binary choices indicating which

jokes participants thought their targets would like more. Participants were asked, “Which of these two test jokes do you think [Person X] liked more?” For each of these five choices, participants also rated their confidence in response to the question “How confident are you in this choice?” Participants rated their confidence on a scale from 50 to 100 to indicate the likelihood that their choice was correct. Finally, at the end of the study, participants gave their own rating to each joke 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”)

**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: Screenshot from Study 1B**

On the left are sample jokes, with the target's ratings. On the right are the test jokes. Below, participants choose (and rate their confidence) which test joke they think the target rated as funnier. We have removed the text of the jokes here, though **a list of all jokes used in all studies are provided online.**

*Here are some jokes Person A rated...*

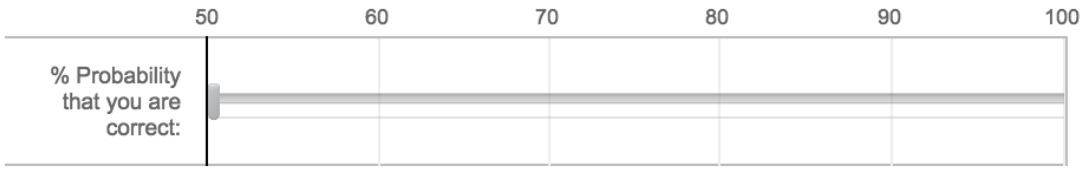
| Sample Joke 1 | Test Joke 1 |
|---|---|
| [text of sample joke 1]<br>Person A rated this joke 4.4 | [text of test joke 1] |
| **Sample Joke 2** | |
| [text of sample joke 2]<br>Person A rated this joke 0 | |
| **Sample Joke 3** | **Test Joke 2** |
| [text of sample joke 3]<br>Person A rated this joke 0.9 | [text of test joke 2] |
| **Sample Joke 4** | |
| [text of sample joke 4]<br>Person A rated this joke 4.4 | |

*Which of these two test jokes do you think Person A liked more?*

Test Joke 1 ○ Test Joke 2 ○

*How confident are you in this choice?*

(remember: 50% is a random coinflip, and 100% is absolute certainty)



## Appendix B: 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 C: 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 D: 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”.