Check for updates

Received: 22 March 2018 | Revised: 8 January 2019 | Accepted: 8 January 2019

DOI: 10.1002/bdm.2118

RESEARCH ARTICLE

WILEY

Making sense of recommendations

Michael Yeomans[1] | Anuj Shah[2] | Sendhil Mullainathan[2] | Jon Kleinberg[3]

[1] Negotiations, Organizations, and Markets, Harvard Business School, Boston, MA, USA

[2] Behavioral Science, University of Chicago Booth School of Business, Chicago, IL, USA

[3] Department of Computer Science, Cornell University, Ithaca, NY, USA

Correspondence
Michael Yeomans, Negotiations, Organizations, and Markets, Harvard Business School, Boston, MA.
Email: myeomans@hbs.edu

Abstract

Computer algorithms are increasingly being used to predict people's preferences and make recommendations. Although people frequently encounter these algorithms because they are cheap to scale, we do not know how they compare to human judgment. Here, we compare computer recommender systems to human recommenders in a domain that affords humans many advantages: 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

algorithms, decision-making, machine learning, recommendations

1 | INTRODUCTION

Computer algorithms can make all manner of predictions. And over the past two decades, the scope of these predictions has broadened significantly. Most notably, algorithms now predict people's preferences across a variety of domains. They make recommendations about which movies and books people will like, which products they should buy, and which restaurants they should visit (Adomavicius & 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 individual decision-making (Adomavicius, Bockstedt, Curley, & Zhang, 2018) and 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 seek out the opinions of other people: 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 Company, 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?

Simple recommender systems, such as "collaborative filtering" algorithms, do not need any information about the *content* of what they are recommending (systems *can* incorporate content, but do not need to; Lops, De Gemmis, & Semeraro, 2011; Gomez-Uribe & Hunt, 2016). Instead, they learn which experiences are statistically similar (Breese, Heckerman, & Kadie, 1998; Koren & Bell, 2011). For example, 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 collect some ratings for each new users and recommend a movie whose ratings correlate strongly with the movies that user rated highly. The same algorithm could make recommendations for any similar data set, regardless of the domain (e.g., restaurants, books, and cars). Clearly, such general-purpose algorithms are convenient. But are they accurate?

 © 2019 John Wiley & Sons, Ltd.

Recommender systems share some clear advantages over human recommenders. These algorithms have perfect memory and execute consistent computation. However, people also seem to have many advantages, particularly for predicting subjective preferences. They often have rich information about their recipients—we often seek and receive recommendations from the people closest to us. Furthermore, people have direct experience with the content of the items (e.g., they know the plot of a movie or the melody of a song). By contrast, recommender systems operate in the dark. They have limited information about the unique tastes of the recipient and little understanding of what they are recommending.

In this paper, we present the first rigorous test of whether recommender systems or humans more accurately predict preferences. In our studies, we collect recommendations from humans and algorithms, both trying to predict which jokes a given person will like. We then compare these recommendations to the jokes those 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 prefer the recommended items (Krishnan, Narayanashetty, Nathan, Davies, & Konstan, 2008; Sharma & Cosley, 2015; Sinha & Swearingen, 2001).

We chose joke recommendations as our test domain for several reasons. First, humor shares many distinctive features of other preference domains. As with other matters of taste, humor is highly subjective. As in other domains, there are genres of humor and heterogeneity in people's preferences for those genres (some people appreciate political humor, others find ribald jokes funny, etc.). Moreover, it is a domain with which most people have significant experience. Humor plays a role in many social interactions in our personal (Bressler, Martin, & Balshine, 2006; Curry & Dunbar, 2013; Martin, Puhlik-Doris, Larsen, Gray, & Weir, 2003) and professional lives (Bitterly, Brooks, & Schweitzer, 2017; Romero & Cruthirds, 2006; Warren, Barsky, Mcgraw, & MacInnis, 2018; Warren & McGraw, 2016). The demands of these interactions will often require people to predict their listeners' reactions to jokes. In that sense, this domain is a familiar perspective-taking task, where common human intuition may be at an advantage.

A practical reason for choosing this domain is that it is easy to read and respond to several jokes in one sitting. Doing so with other preference domains (e.g., movies, books, restaurants, and dating partners) is less practical—instead, previous research in this domain has relied on proxy measures of recommender performance—predicted liking of novel items, or else remembered liking of past items (Krishnan et al., 2008; Sharma & Cosley, 2015; Sinha & Swearingen, 2001). Jokes are an ideal domain for having people experience a common set of novel instances and make several ratings and recommendations in a relatively short time frame.

For our studies, we created a simple recommender system solely on the most basic principles of collaborative filtering— that is, by only using people's preference ratings for individual items. More sophisticated systems could combine this kind of data with other inputs that capture humans' domain knowledge. For example, a more complex model could use human-generated labels of the items' features (e.g., genre, author, and year) along with the ratings data (Gomez-Uribe & Hunt, 2016; Lops et al., 2011). But as a conservative test, our system does not use this kind of content-based input—it does not draw on humans' domain expertise (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 relies only on correlations between ratings of jokes. Our first finding (Studies 1A-1B) shows that, despite their disadvantages, recommender systems outperform human recommenders, even in a 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, in which the primary driver of aversion is algorithmic errors (for a review, see Dietvorst, Simmons, & Massey, 2015). This might also explain some of the aversion to recommender systems, but our final two studies suggest that there is an additional factor that affects trust, even when the error is held constant.

Prior research suggests that people want recommender systems to provide recommendations that they can make sense of (Herlocker, Konstan, Terveen, & Riedl, 2004; McNee, Riedl, & Konstan, 2006). But simple collaborative filtering recommender systems might be particularly difficult to understand because they do not rely on content or on a human to articulate a model of what features guide preferences within a domain. 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, even when people know something is a machine (Fernbach, Sloman, Louis, & Shube, 2012; Keil, 2003; Rozenblit & Keil, 2002; Tintarev & Masthoff, 2011). Accordingly, 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 do not rely on content or human experts to identify the features that are relevant for making predictions. Finally, we show that aversion to algorithms does not merely stem from the content of the recommendations. Instead, it also depends on whether people feel like they can understand the algorithms.

For all studies, sample sizes were set in advance, 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 S1), (2) they did not complete the study, (3) they did not follow instructions, (4) they failed a manipulation check (see Appendix S2), or (5) they rated all jokes as equally funny. The full set of all measures from every study (including

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

exclusion criteria) is described in Supporting Information, and all data and analysis code in R (R Core Team, 2018) are posted on https://osf.io/8nbr2/.

2 | STUDY 1

2.1 | 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 we confirmed that most pairs knew each other very well. Thirty-six percent of pairs were married to one another, another 18% were unmarried couples, and another 30% were immediate family. By another measure of familiarity, 71% of participants had known each other for longer than 5 years and 84% longer than 2 years.

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, Roeder, Gupta, & Perkins, 2001) presented in a random order; all participants saw the same jokes. Participants rated each joke on a slider scale from −10 (*not funny at all*) to +10 (*extremely funny*). The slider was centered at 0 to begin, but participants were required to make a response for each question to advance.

Next, participants completed the recommendation phase of the experiment (see Appendix S3 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"), using the exact same slider scale as their partner. Thus, we had targets' *actual* ratings of the test jokes and recommenders' *predictions* about the targets' ratings of the test jokes (note that recommenders never saw targets' ratings of the test jokes). Finally, we asked participants some basic demographic questions (including how they knew one another) and some exploratory measures concerning their lay beliefs about algorithms.

Our algorithm runs a series of ordinary least squares 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. This simple, easy-to-implement linear regression model will be used as the main comparison against human recommenders throughout our experiments in this paper. However, we will also discuss how other types of collaborative filtering algorithms might perform in the robustness section below.

The algorithm was trained on a dataset of joke ratings from 454 participants culled from this and other studies we had run using this set of 12 jokes. 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 (and when the algorithm predicted a value outside this range, the prediction was coerced to the closest boundary). 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. Additionally, our results are substantively identical if we define prediction error as mean absolute deviation (i.e., L1-distance rather than L2-distance), though we omit those analyses here for brevity.

2.2 | Study 1A 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 [4.095, 4.481]) than human recommenders (RMSE = 5.586, bootstrapped 95% CI [5.314, 5.890]), $t(121) = 6.90$, $P < 0.001$. This was even true when we limited our comparisons to focus specifically on people who were married to their partner (human RMSE = 5.982, bootstrapped 95% CI [5.517, 6.518]; machine RMSE = 4.515, bootstrapped 95% CI [4.192, 4.872]), $t(43) = 4.29$, $P < 0.001$. And it was true for people who had known their partner for more than five years (human RMSE = 5.850, bootstrapped 95% CI [5.517, 6.222]; machine RMSE = 4.493, bootstrapped 95% CI [4.271, 4.740]), $t(86) = 5.21$, $P < 0.001$.

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, Wilson, Lee, & Gilbert, 2015; Gilbert, Killingsworth, Eyre, & Wilson, 2009; Hoch, 1987) suggests that people can make 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 predict that the target would give a joke the same rating as themselves. In our study, human recommenders outperformed mere surrogation (RMSE = 6.495, bootstrapped 95% CI [6.256, 6.749]), $t(121) = 5.81$, $P < 0.001$. The fact that our participants outperformed this easy benchmark suggests that they were invested in the task. But they could not match the performance of the recommender system.

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To our knowledge, this is the first experiment that compares people's preference ratings of novel 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 perhaps knowing each other is what interfered with participants' accuracy. Prior work has found that predictions of other people's preferences can sometimes be *clouded* by knowing each other well, if that extra knowledge distracts judges from the most diagnostic features (Davis, Hoch, & Ragsdale, 1986; Hall, Ariss, & Todorov, 2007; 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.

2.3 | 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 100 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 give a basic sense to participants of how to calibrate their use of the scale. 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., a linear regression recommender system with leave-one-out cross-validation), and the training set for the algorithm was a dataset that included ratings for all 30 jokes from 929 participants in previous studies.

2.4 | Study 1B Results

Once again, machine recommenders outperformed human recommenders. Specifically, the recommender system was more accurate (RMSE = 4.645, bootstrapped 95% CI [4.562, 4.728]) than humans within the "no information" condition (RMSE = 6.087, bootstrapped 95% CI [5.901, 6.282]), within-subjects $t(247) = 15.22$, $P < 0.001$, as well as within the "base rate" condition (RMSE = 5.873, bootstrapped 95% CI [5.710, 6.044]), within-subjects $t(271) = 12.87$, $P < 0.001$. Moreover, in a between-subjects comparison across the two conditions, humans recommenders were only slightly more accurate when they were given base rate information $t(518) = 1.39$, $P = 0.166$. This suggests that the recommender system outperforms humans even when they have some extra information about how people use the scale. It is possible that this was not the most useful information—we could have given participants the raw distribution, or summary statistics about the variation for each joke. In Study 2, we give participants even more information—the machine prediction itself.

2.5 | 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.

However, there may be three 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). 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 [4.095, 4.481]) outperformed the simple average (RMSE = 4.606, bootstrapped 95% CI [4.438, 4.786]), $t(121) = 3.39$, $P < 0.001$. This was also true for Study 1B (machine: RMSE = 4.645, bootstrapped 95% CI [4.562, 4.728]; average: RMSE = 4.966, bootstrapped 95% CI [4.918, 5.016]), $t(519) = 10.2$, $P < 0.001$. These results suggest that there is reasonable heterogeneity across people's preferences in this domain, and the recommender system is able to perform well by detecting these idiosyncrasies.

A second concern might be that our results are specific to our choice of algorithm. Indeed, there are many approaches to collaborative filtering, which can be broadly grouped in terms of two approaches. Item-based collaborative filtering ("IBCF") uses regression-like

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

approaches to predict the ratings of the target item based on the ratings of the sample items (Sarwar et al., 2001). The OLS algorithm used throughout this paper was an example of this. More often these models include some kind of regularization to bias the model away from overfitting, like LASSO (Tibshirani, 1996). Sophisticated models also can include more complex features calculated from the sample ratings, such as nonlinear terms and interactions, or latent factor models. By contrast, user-based collaborative filtering ("UBCF") algorithms gather users from the dataset that have similar preferences on the sample ratings and make predictions by taking a weighted average of these users' ratings on the target item (Breese et al., 1998).

In Appendix S4, we report accuracy estimates for these algorithms across the datasets from Studies 1A and 1B. In general, we found that although there are some subtle differences between approaches, they all are comfortably more accurate than human recommenders. And in keeping with more sophisticated recommender systems, the most accurate was an ensemble model that combined the predictions from a variety of simpler models (Study 1A: RMSE = 4.266, bootstrapped 95% CI [4.077, 4.472]; Study 1B: RMSE = 4.638, bootstrapped 95% CI [4.556, 4.722]). Therefore, our results appear to be robust to the choice of algorithm.

A final 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, whereas 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 using a non-parametric measure of performance 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%]), $t(121) = 2.65$, $P = 0.009$. 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%]), $t(519) = 4.91$, $P < 0.001$.

Finally, in another study (reported in full in Appendix S5), 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%]), $t(196) = 3.15$, $P = 0.002$. These results provide further evidence that machines did not outperform humans merely due to an artifact of the recommendation procedure.

These initial studies provide the first rigorous evidence that collaborative filtering recommender systems can outperform human recommenders. And they do so without having a model of which features of jokes are most predictive of how funny they will be. Put another way, the exact same statistical procedure could be applied to any database of ratings, whether they are of restaurants, books, or cars. The recommender system could make recommendations in any of these domains without knowing what made a restaurant, book, or car enjoyable. But because these recommender systems lack a model of *why* jokes are funny, people might be reluctant to use these machine recommendations as decision aids. In our remaining studies, we continue to develop evidence that recommender systems outperform human recommenders, but our focus now shifts to a related question: Are people reluctant to substitute machine recommendations for human judgment? And if so, why? Figure 1.

3 | STUDY 2

3.1 | 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 2 × 2 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).

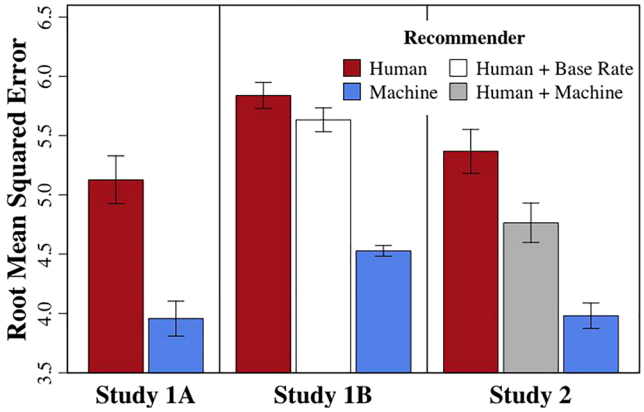

FIGURE 1 Accuracy results from Studies 1 and 2 comparing human recommendations and machine recommendations (error bars represent standard error of the mean) [Colour figure can be viewed at wileyonlinelibrary.com]

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To generate just-in-time predictions from the machine during the current study, we needed to build the recommender system prior to conducting the study. This recommender system was developed using the exact same training data and estimation algorithm from Study 1A. It was implemented as a web service in Python to provide personalized recommendations based on people's sample joke ratings. Unlike Study 1A, we did not need to use cross-validation to estimate held-out accuracy, because the predictions in this study were being made for participants whose data we did not have yet.

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 a museum goer 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).

3.2 | Results

Regarding accuracy, recommender systems (RMSE = 4.273, bootstrapped 95% CI [4.128, 4.291]) once again outperformed humans (RMSE = 5.387, bootstrapped 95% CI [5.199, 5.583], $t(209) = 10.06$, $P < 0.001$). And the humans did not perform any better for close others (RMSE = 5.386, bootstrapped 95% CI [5.126, 5.673]) or for strangers (RMSE = 5.387, bootstrapped 95% CI [5.134, 5.663]), $t(208) = 0.25$, $P = 0.802$.

Despite the fact that machines were more accurate than humans, people did not entirely rely on machine recommendations to guide their judgments. People did improve somewhat when given the machine predictions (RMSE = 5.056, bootstrapped 95% CI [4.820, 5.314]) compared with those without it (RMSE = 5.692, bootstrapped 95% CI [5.420, 5.991]), $t(208) = 2.42$, $P = 0.017$. But the humans with the recommender system still performed worse than the recommender system on its own (RMSE = 4.110, bootstrapped 95% CI [3.912, 4.323]), $t(103) = 6.61$, $P < 0.001$. These data suggest that people did not ignore the machine recommendation. But they also did not perform as well as they could have, had they been willing to put more trust in the recommender system.

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). The results above suggest that people may be similarly reluctant to rely on recommender systems. It is worth noting that this reluctance may be quite rational. Given that people have more experience with their own recommendation process (and that of other people), some initial reluctance to embrace a new recommendation technology seems reasonable. People might have wanted more information about the algorithm (such as feedback on whether its predictions were accurate). However, our findings do suggest that human judgment could be improved if people more fully incorporated the recommender system's predictions into their own.

Of course, people rarely use recommender systems to help them *make* recommendations for other people. Instead, people most often interact with recommender systems when they are *receiving* recommendations for themselves, and in those cases, they will have better information about the whether a recommender system matched their own preferences. Are people averse to *receiving* machine recommendations? We address this question in the following studies Figure 2.

4 | STUDY 3

If people are averse to machine recommendations, then this could be due to two factors. First, machine recommenders select different contents (i.e., which jokes they recommend). Second, machine recommenders use a different recommendation process than do humans (or, at least, people *believe* that the recommender systems follow a different process). 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).

4.1 | 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.

4.1.1 | 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

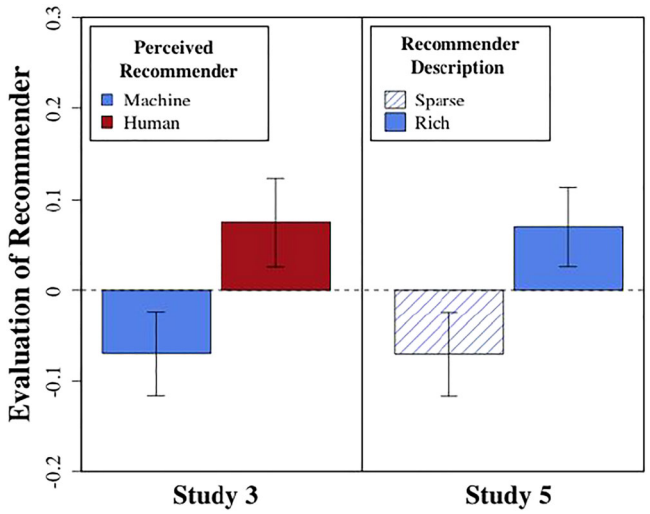

FIGURE 2 Recipients' evaluations of the recommenders' ability from Studies 3 and 5, based on perceived recommendation source and recommender description, respectively (error bars represent standard error of the mean) [Colour figure can be viewed at wileyonlinelibrary.com]

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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, and Sample Joke 3: 7.5). Recommenders then picked three test jokes (from a menu of 10) 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 as inputs. To make predictions, the algorithm used the same linear regression approach as in the studies above, and the ratings data were the same as in Study 1B—929 participants who had rated all 13 jokes used in the current study.

4.1.2 | 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 2 × 2 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 because 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 exact calculations of the algorithms involved.

Participants first rated three sample jokes and 10 test jokes. They then waited 20 s 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.

4.2 | Results

4.2.1 | 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]), $t(885) = 3.01$, $P = 0.003$.

4.2.2 | Aversion to recommender systems

Next, we compared how participants rated the *recommenders*. As planned, there was high internal consistency among the three evaluation questions (Cronbach's $\alpha = 0.95$), so we standardized and combined responses into a single "preference index" (and all our results are substantively identical if we compute a simple average of these scales instead, or analyze scales individually). A 2 × 2 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, $F(1, 882) = 1.8$, $P = 0.178$.

These results demonstrate another dimension of aversion to recommender systems. In Study 2, people did not make full use of recommender systems when recommending for others, and here, we see they are also reluctant to receiving recommendations from recommender systems. 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—the average funniness of the three jokes they saw—was meaningfully correlated with the preference index, $r = 0.59$, $t(884) = 22.0$, $P < 0.001$. Accordingly, we conducted a robustness check using a multiple linear regression, which included both variables as predictors. When controlling for the effect of recommender accuracy, the regression model estimated a larger effect of perceived source, $\beta = 0.183$, $SE = 0.054$, $t(883) = 3.4$, $P < 0.001$. This model also benchmarked the relative effect sizes of actual accuracy and perceived recommendation source—we estimate that the implicit penalty against the machine was equivalent to a difference in accuracy of 0.31 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,

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2000; Tintarev & Masthoff, 2011), and perhaps 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 Figure 3.

5 | STUDY 4

5.1 | Methods

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

The study was identical to Study 3, with two exceptions. First, as a measure of our proposed mediator, participants were asked to rate how easy it was to understand the recommendation process. They did this by reporting 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." This choice question was always asked last, after the understanding questions, and the order of the two options was counterbalanced.

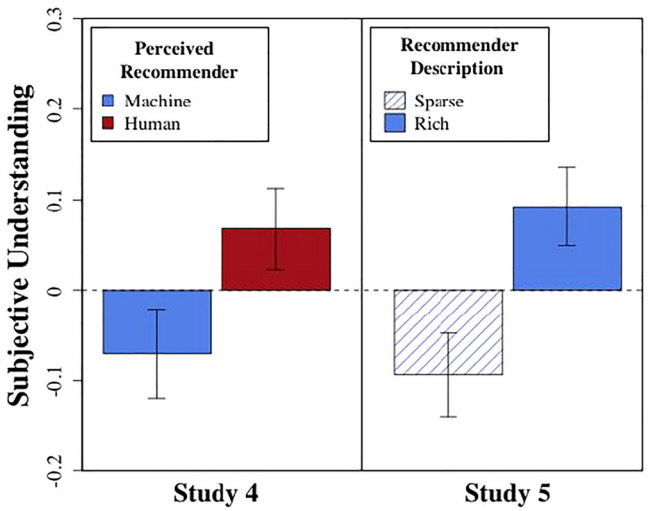

FIGURE 3 Recipients' rated understanding of the recommenders from Studies 4 and 5, based on perceived recommendation source and recommender description, respectively (error bars represent standard error of the mean) [Colour figure can be viewed at wileyonlinelibrary.com]

5.2 | 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]), $t(898) = 4.15$, $P < 0.001$.

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. This is an interesting artifact of our experimental design, perhaps a result of novelty seeking. But more relevant to our hypotheses, we found that 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%]), $\chi^2(1, N = 899) = 12.84$, $P < 0.001$. Put simply, a majority of participants preferred human recommenders (M = 54.8%, 95% CI [51.6%, 58.1%]), $\chi^2(1, N = 899) = 8.42$, $P = 0.004$.

The subjective understanding ratings were combined in a single index (Cronbach's $\alpha = 0.82$), and we again confirm our results are consistent if we instead compute a simple average or analyze individual items. 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]), $t(897) = 2.07$, $P = 0.038$. 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, $\beta = 0.60$, $SE = 0.09$), $z(897) = 7.01$, $P < 0.001$. And this relationship was attenuated when participants thought their recommender was human, $\beta = 0.43$, $SE = 0.11$, $z(457) = 3.83$, $P < 0.001$; interaction term: $\beta = -0.39$, $SE = 0.18$, $z(895) = 2.19$, $P = 0.028$. Like in Study 3, this estimated relationship is stronger when we control for the accuracy of the recommender itself (interaction term: $\beta = -0.57$, $SE = 0.20$), $z(893) = 2.86$, $P = 0.004$. Furthermore, this effect was significantly mediated by subjective understanding (bootstrapped indirect effect: $M = 0.015$, 95% CI [0.030, 0.00], $P = 0.039$; Tingley, Yamamoto, Hirose, Keele, & Imai, 2014). In other words, this suggests that people are averse to using recommender systems in part 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

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

recommenders when they do not feel like they understand *how* they make recommendations. And this subjective understanding seems to be malleable, independent of the performance of the recommender itself. Would making machine recommendations easier to understand increase how much people like those recommenders? The final study addresses this possibility.

6 | STUDY 5

6.1 | Methods

One thousand and fourteen participants from MTurk completed our study. Twenty-four participants failed the manipulation check and four 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. Because the chosen jokes were rated after the manipulation, this allowed us to test whether our manipulation affected the subjective quality of the recommended items themselves. Second, all recommendations were generated by a recommender system that used the exact (i.e., unrounded) 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 was 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 S6 for exact stimuli).

6.2 | 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]), $t(984) = 2.93$, $P = 0.003$. This confirmed that our manipulation had its intended effect. One concern might be that our manipulation not only affected subjective understanding but also affected people's perceptions of recommender accuracy. That is, perhaps the rich explanation made it seem like the algorithm had chosen better jokes. If this were true, then we might expect to find that people rated the recommended jokes more favorably in the rich explanation condition. However, there was no difference in the average ratings of the three chosen jokes in the sparse condition (M = 3.52, 95% CI [3.18, 3.86]) and the rich condition (M = 3.55, 95% CI [3.23, 3.87]), $t(984) = 0.13$, $P = 0.898$. This suggests that our manipulation affected subjective understanding, but not necessarily subjective accuracy.

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]), $t(984) = 2.20$, $P = 0.028$. As in previous studies, this effect was stronger when we controlled for accuracy of the recommender (β = 0.13, SE = 0.04), $z(983) = 3.09$, $P = 0.002$. Furthermore, this effect was significantly mediated by subjective understanding (bootstrapped indirect effect: M = 0.104, 95% CI [0.035, 0.18], $P = 0.003$). In other words, rich explanations of the recommender system increased participants' understanding of the recommendation process, and this in turn improved their beliefs about the quality of the recommender system's performance.

7 | GENERAL DISCUSSION

The ubiquity of recommender systems raises a familiar question: How do algorithmic predictions compare to human judgment? How do algorithmic predictions compare to human judgment? Here, we find a familiar answer—recommender systems can outperform human recommenders, even when those humans are making recommendations for friends and family.

These results build on a large body of work comparing human and algorithmic judgment. Early research discovered that algorithms could produce more accurate judgments than humans could, when they are given the same features or cues for making judgments (Dawes, 1979; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954). Those algorithms often improved on human judgment by making it more consistent. 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 and 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). By contrast, collaborative filtering is designed to perform well even when this expertise is not used to curate the available features. Recommender systems can excel even with 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. To be sure, there are many commonalities between these two kinds of models—both succeed by consistently applying prediction algorithms to a database—and in practice, these approaches are often intricately combined. But our results add another datapoint suggesting that perhaps algorithms do not require the special expertise of humans to outperform them

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

It is possible that humans might still outperform recommender systems in other domains. Our results suggest that even for something like humor, which seems uniquely human, recommender systems can excel. But it is still worthwhile for future research to conduct similar tests in other preference domains. Recommender systems also suffer from start-up problems in new applications, when the database of previous ratings is small, or sparse. In practice, the accuracy of recommender systems may also be bounded by their strategies to collect user ratings, suggesting questions for future work on the behavioral aspects of review elicitation (Avery, Resnick, & Zeckhauser, 1999; Berger, 2014; Burtch, Hong, Bapna, & Griskevicius, 2017; Yeomans, 2018).

Our findings also highlight a familiar theme from prior work comparing algorithmic judgment to human judgment. People are averse to relying on recommender systems. People do not sufficiently use input from these systems and they prefer to receive recommendations from humans. For recommendation recipients, this is not irrational. For one, it can take time (and a rich track record of success) for new technology to earn a reputation for performance in the public mind. More broadly, people can have a direct preference for process, for example, by privileging predictions made with domain-specific content rather than mere ratings data (Gilbert et al., 2009). And preferences for algorithmic and human judgment also vary across domains. In particular, domains that are more subjective—like jokes—may be the most likely to elicit algorithmic aversion (Castelo, Lehmann, & Bos, 2019; Logg, 2018). But despite these reservations, we still think that people may be missing opportunities to get more accurate recommendations because of their reluctance to rely on recommender systems.

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 does seem to draw from other factors that can be independent of accuracy. In particular, we find that one reason people seem averse to recommender systems is because they do not understand the recommendation process, and they 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 (Fernbach et al., 2012; 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 accuracy is not the sole determinant of subjective preferences. In some sense, if the next "Netflix Challenge" focused on facilitating algorithmic sensemaking, it might provide further encouragement for people to 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). In some cases, it may also be that simply allowing people more experience with recommender systems will increase feelings of understanding over time. 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 without the special expertise of a human. But accuracy alone cannot reduce aversion to algorithms—they need to be understood as well.

CONFLICT OF INTEREST

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 https://osf.io/8nbr2/.

ORCID

Michael Yeomans https://orcid.org/0000-0001-5651-5087

REFERENCES

Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers' willingness to pay. *Information Systems Research* In press, *29*, 84–102. https://doi.org/10.1287/isre.2017.0703

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. https://doi.org/10.1109/TKDE.2005.99

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet Recommendation Systems. *Journal of Marketing Research*, *37*(3), 363–375. https://doi.org/10.1509/jmkr.37.3.363.18779

Avery, C., Resnick, P., & Zeckhauser, R. (1999). The market for evaluations. *American Economic Review*, *89*(3), 564–584. https://doi.org/10.1257/aer.89.3.564

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, *9*(2), 75–79. https://doi.org/10.1145/1345448.1345465

Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, *24*(4), 586–607. https://doi.org/10.1016/j.jcps.2014.05.002

Bitterly, T. B., Brooks, A. W., & Schweitzer, M. E. (2017). Risky business: When humor increases and decreases status. *Journal of Personality*

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and Social Psychology, *112*(3), 431–455. https://doi.org/10.1037/pspi0000079

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research*, *45*(1), 77–93. https://doi.org/10.1509/jmkr.45.1.77

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. https://doi.org/10.1016/j.obhdp.2006.07.001

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.

Bressler, E. R., Martin, R. A., & Balshine, S. (2006). Production and appreciation of humor as sexually selected traits. *Evolution and Human Behavior*, *27*(2), 121–130. https://doi.org/10.1016/j.evolhumbehav.2005.09.001

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. https://doi.org/10.1287/mnsc.1110.1371

Buell, R. W., & Norton, M. I. (2011). The labor illusion: How operational transparency increases perceived value. *Management Science*, *57*(9), 1564–1579. https://doi.org/10.1287/mnsc.1110.1376

Burtch, G., Hong, Y., Bapna, R., & Griskevicius, V. (2017). Stimulating online reviews by combining financial incentives and social norms. *Management Science*, *64*(5), 2065–2082.

Castelo, N., Lehmann, D., & Bos, M. (2019). Consumers' use of algorithms. *Working paper*.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Curry, O. S., & Dunbar, R. I. (2013). Sharing a joke: The effects of a similar sense of humor on affiliation and altruism. *Evolution and Human Behavior*, *34*(2), 125–129. https://doi.org/10.1016/j.evolhumbehav.2012.11.003

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. https://doi.org/10.1086/209045

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

De, P., Hu, Y., & Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science*, *56*(11), 1930–1945. https://doi.org/10.1287/mnsc.1100.1233

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), 114–126. https://doi.org/10.1037/xge0000033

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*, *64*, 1155–1170.

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. https://doi.org/10.1016/j.obhdp.2015.03.005

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2012). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*(5), 1115–1131.

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature*, *75*(7), 450–451. https://doi.org/10.1038/075450a0

Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., & Wilson, T. D. (2009). The surprising power of neighborly advice. *Science*, *323*(5921), 1617–1619. https://doi.org/10.1126/science.1166632

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, *4*(2), 133–151. https://doi.org/10.1023/A:1011419012209

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 Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes*, *103*(2), 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003

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

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. https://doi.org/10.1145/963770.963772

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology*, *53*(2), 221–234. https://doi.org/10.1037/0022-3514.53.2.221

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373. https://doi.org/10.1016/S1364-6613(03)00158-X

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

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., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

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. https://doi.org/10.1086/508523

Logg, J. (2018). When do people rely on algorithms? *Working Paper*.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook* (pp. 73–105). US: Springer.

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. https://doi.org/10.1287/mnsc.1090.1031

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, *37*(1), 48–75. https://doi.org/10.1016/S0092-6566(02)00534-2

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*. US: University of Minnesota Press.

Nielsen Company (2015). *Global trust in advertising survey. September 2015*. www.nielsen.com

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–265. https://doi.org/10.1037/0033-295X.84.3.231

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58. https://doi.org/10.1145/245108.245121

Romero, E. J., & Cruthirds, K. W. (2006). The use of humor in the workplace. *The Academy of Management Perspectives*, *20*(2), 58–69. https://doi.org/10.5465/amp.2006.20591005

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

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*.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B: Methodological*, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38.

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

Warren, C., Barsky, A., Mcgraw, A. P., & MacInnis, D. (2018). Humor, comedy, and consumer behavior. *Journal of Consumer Research*. In press, https://doi.org/10.1093/jcr/ucy015

Warren, C., & McGraw, A. P. (2016). When does humorous marketing hurt brands? *Journal of Marketing Behavior*, *2*(1), 39–67. https://doi.org/10.1561/107.00000027

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388. https://doi.org/10.1016/j.tics.2010.05.006

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. https://doi.org/10.1016/j.jesp.2014.01.005

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yeomans, M. (2018). Some hedonic consequences of perspective-taking in recommending. *Journal of Consumer Psychology*, in press, *29*, 22–38.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Yeomans M, Shah A, Mullainathan S, Kleinberg J. Making sense of recommendations. *J Behav Dec Making*. 2019;32:403–414. https://doi.org/10.1002/bdm.2118

10990771, 2019, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2118 by Massachusetts Institute of Technolo, Wiley Online Library on [23/07/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License