Received: 17 February 2021 | Accepted: 16 March 2021

DOI: 10.1002/aaai.12055

SPECIAL TOPIC ARTICLE

Recommender systems, ground truth, and preference pollution

Gediminas Adomavicius[1] | **Jesse C. Bockstedt**[2] | **Shawn P. Curley**[1] | **Jingjing Zhang**[3]

[1]Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota, USA

[2]Information Systems and Operations Management, Goizueta Business School, Emory University, Atlanta, Georgia, USA

[3]Operations and Decision Technologies, Kelley School of Business, Indiana University, Bloomington, Indiana, USA

Correspondence
Jingjing Zhang, Operations and Decision Technologies, Kelley School of Business, Indiana University, 1309 East Tenth Street, HH4143, Bloomington, IN 47405, USA.
Email: jjzhang@indiana.edu

Abstract

Interactions between individuals and recommender systems can be viewed as a continuous feedback loop, consisting of pre-consumption and post-consumption phases. Pre-consumption, systems provide recommendations that are typically based on predictions of user preferences. They represent a valuable service for both providers and users as decision aids. After item consumption, the user provides post-consumption feedback (e.g., a preference rating) to the system, often used to improve the system's subsequent recommendations, completing the feedback loop. There is a growing understanding that this feedback loop can be a significant source of unintended consequences, introducing decision-making biases that can affect the quality of the "ground truth" preference data, which serves as the key input to modern recommender systems. This paper highlights two forms of bias that recommender systems inherently inflict on the "ground truth" preference data collected from users after item consumption: non-representativeness of such preference data and so-called "preference pollution," which denotes an unintended relationship between system recommendations and the user's post-consumption preference ratings. We provide an overview of these issues and their importance for the design and application of next-generation recommendation systems, including directions for future research.

INTRODUCTION

Personalized recommender systems are an accepted and valued component of many online experiences, including retail shopping (e.g., on Amazon), movie watching (e.g., Netflix), and music listening (e.g., Spotify, Pandora). Due to the rapidly growing ubiquity of recommender systems, understanding the impact of online personalization across a wide variety of dimensions has become an increasingly important research paradigm (Baeza-Yates 2020; Zanker et al. 2019). In this paper, we discuss how user-recommender interactions inherently (and disadvantageously) affect the quality of the "ground truth" preference data for any recommender system. This is an important issue, because preference data serves as the key input to recommender systems, for example, one of the goals of many recommender systems is to predict user preferences for the yet-unconsumed items based on known preference data. This is by no means the only goal, as recommender systems' algorithms have been designed

to incorporate additional useful considerations, such as diversity, novelty, fairness, and value-awareness. However, for successful personalization applications, predicting or estimating user preferences is a necessary, crucial component. Thus, many approaches to recommender systems use predictive modeling techniques based on machine learning methodology.

Training data for modern recommender systems, especially for collaborative filtering approaches that represent the most popular and widely used recommendation techniques, is typically data on user-item interactions. Such user-item interaction data can represent *implicit* preference data (views, clicks, purchases) or *explicit* preference data (self-reported user preference ratings of items). Both types of data can be (and have been) used by recommendation algorithms, although each has its benefits and limitations. For example, implicit data is often much more readily available (just by observing user activities); in contrast, explicit data requires deliberately prompting the user to respond to an interface for data collection in the form of ratings and, thus, could be seen as more intrusive or requiring more effort. On the other hand, implicit data may represent a weaker preference signal. The fact that a user browsed (or even purchased) a product on Amazon may indicate some level of preference/interest for that item. However, the number of 1-star and 2-star reviews for confirmed purchases emphasizes the fact that browsing and even purchasing are typically pre-consumption activities, that is, the user has not yet had a chance to consume the item and formulate an informed preference. Explicit rating data, however, is typically provided by the users after experiencing or consuming the item and, thus, provides a preference signal that could be more informative for learning by a recommendation algorithm.

Consequently, as many recommender systems are based on predictive modeling, a critical, necessary component for successful predictive modeling is having advantageous ground truth data on user preferences. However, as with many complex and rapidly developing technologies, online recommender systems may have unintended consequences as side effects. In this paper, we take a closer look at a specific aspect of the recommender systems' impact on users—that is, how recommender systems can bias user preference ratings (instead of merely trying to predict them), compromising their value as ground truth data for learning.

Two general forms of bias are highlighted in the paper, tied to important, desirable characteristics of preference ground truth data in the development and assessment of recommendation systems. There can be numerous reasons for having sub-optimal ground truth in predictive models: cultural (biases, prejudices, etc.), noise, subjective measurements, etc. In this paper, we focus on two important aspects that are directly related to the inherent nature of recommender systems, in general: *preference representativeness* and *preference independence*. The preference representativeness aspect indicates that the user–item interactions in the available ground truth data (explicit or implicit) are representative of (i.e., a random or at least unbiased sample of) the entire space of possible user-item interactions for the given application context. For example, in the general-purpose movie recommender system, if the only available movie ratings were from teenagers about horror movies, this would constitute a non-representative ground truth dataset. The preference independence aspect indicates that the preference information (e.g., user ratings of items) provided for user–item interactions depend only on the users' judgments of the items and not upon the system's predictions against which the ground truth is being compared. In other words, presenting a recommendation to a user should not affect the rating the user provides after she/he has consumed the item.

To understand how the very nature of recommender systems can introduce two general forms of bias related to preference representativeness and preference independence, it is useful to recognize that the interactions between individuals and recommender systems can be viewed as a feedback loop (see Figure 1) and to highlight the distinction between the pre-consumption and post-consumption phases in users' interactions with recommender systems.

Starting at the left side of Figure 1, recommender systems have been explicitly and intentionally designed to affect users' behavior in the *pre-consumption* phase. They are designed to help each user find relevant information (content, products, services, etc.) in the huge sea of available options. Therefore, it is not unexpected to see that recommender systems can directly affect the *item consumption choices* of users; this is indeed a goal of recommender systems. However, this is not to say that all pre-consumption effects of recommender systems are desirable. There has already been a steady stream of studies about various biases in recommender systems (Baeza-Yates 2020; Chen et al. 2020), resulting in a robust discussion of how some inherent aspects of recommender systems algorithms and interfaces—such as popularity bias (Abdollahpouri 2019; Abdollahpouri et al. 2017; Prawesh and Padmanabhan 2014), position bias (Collins et al. 2018; Guo et al. 2019; Hofmann et al. 2014; Wang et al. 2018), and other biases—can result in "filter bubbles," "echo chambers," or outcomes that are biased with respect to some desired "fairness" criteria (e.g., Abdollahpouri et al. 2020b; Ekstrand et al. 2019; Ekstrand et al. 2018; Flaxman et al. 2016; Gao and Shah 2020; Ge et al. 2020; Nguyen et al. 2014; Pariser 2011).

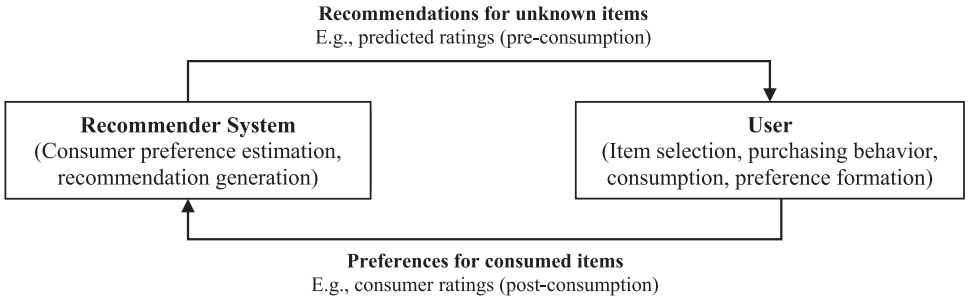

FIGURE 1 Feedback loop in user-recommender interactions (adapted from Adomavicius et al. 2013).

Many different approaches have been proposed to remedy the potential inadequacies of recommender systems in the pre-consumption phase. One common direction is to include additional considerations (i.e., beyond recommendation accuracy) that are valuable to consumers and/or providers (such as diversity, novelty, fairness, budgetary constraints, value awareness) as part of the recommendation process in a given setting (e.g., Adomavicius and Kwon 2012; Azaria et al. 2013; Burke et al. 2018; Jannach and Adomavicius 2017; Jannach and Bauer 2020; Jannach et al. 2012; Lakiotaki et al. 2011). Another (related) direction is to explicitly consider additional stakeholders, whose considerations might be important, to incorporate into recommendation engines. An obvious example would be "societal" considerations, which may go beyond individual consumers' or providers' interests toward avoiding filter bubbles and echo chambers as well as toward incorporating "fairness". The area of *multi-stakeholder* recommendations is a rapidly developing area that looks into some of these issues (e.g., Abdollahpouri et al. 2020a; Sürer et al. 2018; Zheng 2019; Zheng et al. 2019). No matter what algorithm or criteria are employed, however, representativeness of the dataset used to make pre-consumption recommendations is often presumed. In statistics, this form of bias leading to non-representative data is denoted as non-sampling error and is a recognized threat to valid statistical inference. The same concerns apply in the operation of recommender systems, comprising a form of bias that can arise from non-representativeness of the data used as inputs to any recommendation procedure.

While recommendations are explicitly designed to provide value at the pre-consumption stage (i.e., help the users deal with potential information overload by suggesting the most relevant content), they are not presumed to continue providing value or impact after the consumption choice is made and a user experiences the item, that is, at the *post-consumption* phase. After consuming an item, retailers often ask for consumers' reactions in the form of item ratings, for example, using a 5-star scale on Amazon, a 10-star scale on IMDb, or a thumbs-up/down rating on the Netflix streaming service. These preference ratings collected at the post-consumption stage are presumed to be indications of the user's ground truth preference for the recently experienced items. In other words, once the item is consumed, the user's rating should be based on his/her preferences and be independent of the system recommendations that were presented pre-consumption. Under this assumption, the users' preference ratings are routinely used by the same recommender systems as measures of ground truth that can be directly compared to system predictions to assess the accuracy of the recommendations, and to retrain the predictive models and further refine and improve subsequent rating predictions. However, recent research increasingly indicates that the presumption of the independence of users' post-consumption ratings from pre-consumption recommendations does not hold. If the user sees a recommendation prior to making their post-consumption preference rating, their response can be significantly impacted, leading to a pollution of the preference rating. In other words, the observed system recommendations (in the form of system-predicted personalized ratings) continue to affect users' preference ratings, which are then subsequently reported back to the system in the post-consumption phase (Adomavicius et al. 2013; Cosley et al. 2003). For example, users seeing a recommendation that is artificially adjusted upward or downward systematically provide higher or lower preference ratings post consumption, respectively. This phenomenon is not aligned with the normative expectation of the system designers (and likely the users as well) that consumers' preferences, post-consumption, should reflect the consumers' judgments based solely on their experiences of the item and independent of the recommendation. Thus, the post-consumption effects of recommendations on users' preferences is appropriately characterized as a bias relative to this normative expectation. Specifically, we define *preference pollution* as a post-consumption effect of recommendations upon users' stated preferences, in contradiction with normative expectations of no such effect, as commonly held by retailers, system designers, and users.

In the next two sections, we elaborate on each of these two forms of bias, paying particular attention to

preference pollution as an underappreciated area within the recommendation systems literature.

RECOMMENDER SYSTEMS AND PREFERENCE NON-REPRESENTATIVENESS

Traditionally, recommender systems are most useful (and most widely used) in application domains where users are faced with choices from a vast number of alternatives, for example, many movies or TV shows to watch, many books to read, many songs to stream. These also tend to be domains where users exhibit subjective, taste-based preferences; accordingly, personalization technologies have the most impact in such domains. (In contrast, in domains where there is a clear, objective quality to an item, non-personalized recommendations can be made and be extremely successful to all users.) The data used by consumers in taste-based domains is often collected through observation of readily available information. Most users want to consume items that are "good" for them (and avoid items that are "bad" for them) and use any plausibly relevant information to help them with their decisions. In the movie domain, this could mean watching movie trailers, reading reviews of movie critics, listening to friends' opinions, etc. Considering the nature of readily available sources, even absent any recommender systems, the data collected likely will not be "representative," but rather will be substantially skewed towards positive information. In the research literature, this phenomenon is called MNAR (data "missing not at random") (Ishioka 2014; Little and Rubin 2019; Nugroho and Surendro 2019; Santore et al. 2020; Tremblay et al. 2010).

The MNAR phenomenon is significantly exacerbated by the presence of recommender systems in user-item interactions. It is a well-known phenomenon, not limited to recommender systems. Users' consumption choices are significantly affected by what is shown to them. For example, in Internet search, people focus most of their clicks only on a small subset of search results at the top of the provided ranked list. In modern applications (e-commerce, video/music streaming, etc.), what is shown to users is often driven by recommender systems. And, as has been widely understood, MNAR is common in recommender systems, as they produce clearly non-random, non-representative sets of items as recommendations (Baeza-Yates 2018, 2020; Jannach et al. 2015; Mansoury et al. 2020; Marlin et al. 2012). The MNAR issue affects recommender systems regardless of whether they are built on implicit or explicit preference data. For instance, while the presence of implicit data (e.g., views, clicks, and purchases) does represent some indication of a user's preference toward an item, the lack of an interaction (e.g., no click) does not necessarily indicate an item is irrelevant for the user—it may simply be that the user was not aware of the item or chose not to click on the item for any number of plausible reasons. Thus, evaluation measures computed on the observed data may not accurately reflect performance on the complete data (Lim et al. 2015). Generally, recommender systems (by design) will create a systematic bias towards observing more highly rated items (Saito 2020), naturally providing a more skewed representation of users' stated preferences. This, in turn, can be detrimental to the system's performance (Zhang et al. 2020). In summary, recommender systems amplify the MNAR issue and introduce systematic biases into their training data, which subsequently leads to biased predictions of users' preference ratings on unconsumed items.

The MNAR issue in recommender systems has been extensively studied (e.g., Kim and Choi 2014; Marlin et al. 2012; Marlin and Zemel 2009; Saito et al. 2020). Marlin et al. (2012) reported evidence of MNAR in recommender systems based on a large-scale online study conducted at Yahoo! Research. Marlin and Zemel (2009) empirically analyzed the effect of non-random missing data on rating predictions of popular recommendation algorithms. Their analysis shows that recommendation methods that incorporate a non-random missing data model consistently outperform the baseline methods that do not consider MNAR on both rating prediction and item ranking. Using experiments with Yahoo! users, Pradel et al. (2012) showed that ignoring missing items can lead to a dramatically biased evaluation. Meanwhile, considering missing ratings as a form of negative feedback may improve performance, but it is also misleading and can bias evaluation towards models that favor popularity rather than individual user preference.

A variety of techniques have been proposed to address the MNAR problem in recommender systems and to enhance the representativeness of training data accordingly (e.g., Kim and Choi 2014; Marlin and Zemel 2009; Schnabel et al. 2016; Steck 2010; Wang et al. 2019; Yang et al. 2018). For example, Steck (2010) shows that using the top-k hit rate, defined as the fraction of relevant items in the top-k recommendation list, as an accuracy measure for recommendations is better than traditional measures such as root mean square error. Under mild assumptions, the top-k hit rate can be estimated without bias from data even when MNAR is present. Also, Steck (2013) proposes an error-imputation-based approach that computes an imputed error, that is, an estimated value of the prediction error, for each missing rating. The imputed errors are used to recover the prediction errors for missing ratings, or weight observed ratings with the propensities of

being observed (Bertsimas et al. 2017; Wang et al. 2019). Another method is the inverse-propensity-scoring (IPS) approach that inversely weights the prediction error for each observed rating with the propensity of observing that rating. IPS can be applied to both explicit feedback (Schnabel et al. 2016) and implicit feedback (Yang et al. 2018). The propensity score is the probability of each data being observed, and unbiased performance estimation is possible by weighting each data item by the inverse of its propensity. The IPS method is affected by the choice of the propensity estimation model and the high variance problem. To overcome such limitations, recent work further developed an asymmetric tri-training meta-learning method that minimizes the propensity-independent upper bound of the ideal loss function (Saito 2020).

In summary, recommender systems pose an inherent challenge to the *preference representativeness* in the ground truth data. This challenge can be attributed in large part to pre-consumption activities of users, as the consumption choices are (by design) increasingly influenced by recommender systems. However, recommender systems are not explicitly designed to affect the values of the user's post-consumption preference ratings, which are routinely used as "ground truth" for recommendation algorithms. In other words, independence between observed recommendations and users' post-consumption preference ratings is commonly assumed. Recent studies increasingly indicate that this independence does not hold as a general rule, that is, preference pollution is observed. We discuss the issue of preference pollution next.

RECOMMENDER SYSTEMS AND PREFERENCE POLLUTION

Recommender systems represent a highly valuable aspect of user experience during the pre-consumption and/or pre-purchase period, where they help users to find and select relevant items. After consuming items, the user provides preference ratings back to the system, which are new data that can be used to refine the system's subsequent predictions. This process creates a feedback loop. As discussed earlier, a fundamental presumption that is made in recommender systems literature (often implicitly, but sometimes explicitly as well) is that the stated post-consumption preference rating provided by the user represents ground truth. It is assumed to be based on the user's actual experience with the item and independent of the system recommendation presented pre-consumption (setting aside the work that analyzes purposefully malicious user behavior, i.e., attacks on recommender systems (Mobasher et al. 2007)).

But is this assumption valid? With item consumptions increasingly curated by recommender systems, are the post-consumption user ratings truly unaffected by the system ratings? Importantly, recent studies show that interacting with online personalization and recommendation systems can have unintended side effects on user preference ratings and economic behavior—both can be significantly distorted by the system-predicted ratings. This bias, which we defined as *preference pollution*, can have important implications for recommender systems' design and usability. Preference pollution has been underexplored within the recommendation systems literature; thus, in this section we pay extended attention to its ubiquity, possible mitigation strategies, and several important directions for future work. To place the fundamental issue of preference pollution with recommender systems in a broader context, we also present related work in past and present research that studies the effects of decision anchors, persuasiveness, and social influences on user behaviors.

Preference pollution and system-induced biases

The key questions for recommender systems research under consideration in this context are: *Do the system predictions displayed to the user before item consumption unintentionally influence their post-consumption preference ratings? And, if so, what are the implications of this influence?* The normative ideal, and what is generally presumed, is independence, namely that the recommendations do not influence user ratings. Let us return to Figure 1 to clarify the issue. During the pre-consumption and/or pre-purchase period, we want and expect the system ratings to affect user behavior through suggestions of items to consume—this is the value component that the recommendations represent. However, once the user has consumed the item, we want to get an expression of the user's true preference for the item that is unpolluted by the system recommendation. This is especially assumed to hold in the case when the user's preference rating is captured immediately following consumption, when effects due to imprecise memory of a past experience are not operative. When the post-consumption preference rating is impacted by the pre-consumption system recommendation, a preference pollution has occurred. As noted by Cosley et al. (2003), the biases defined by this type of preference pollution can lead to a number of potential problems: they can contaminate the recommender system's inputs, reducing its effectiveness; they can provide a distorted view of the system's performance; and, they can allow agents to manipulate the system to operate in their favor.

The question of possible preference pollution has remote cousins in past and present research that studies the effects

of decision anchors, persuasiveness, and social influences on user behaviors (e.g., see Ariely 2010 or Kahneman 2011 for book-length overviews accessible to a broad audience). However, the personalized recommender systems environment has a number of features that make it unique and call into question the applicability of that research to the present situation. Recent work into the effects of system recommendations on user preference responses has begun to address this question. Furthermore, since the recent studies have primarily used randomized controlled experimental methodologies, we are able to speak to the causality that system recommendations have in producing preference pollution.

In particular, it has been demonstrated that personalized recommendations cause users' post-consumption ratings to become biased in the direction of the recommendation (Adomavicius et al. 2013). Experiments show that when a recommendation is manipulated upward or downward, the consumer's reported preference ratings (for a product they consume) shift significantly in the same direction. On a 1–5 star rating scale, a perturbation of one star in the system prediction leads to a mean user rating shift of about 0.35 stars. This effect is robust across various types of digital good—it has been consistently observed for movies, TV shows, songs, and jokes (Adomavicius et al. 2013, 2018, 2019). The effect occurs for both artificially generated system ratings (irrespective of any actual system behavior), as well as for ratings predicted by a real, validated recommender system that are perturbed either higher or lower. These effects persist when accounting for any individual preference differences. The effect also occurs both for recalled preferences, that is, how much did you like Film X (seen in the past) (Cosley et al. 2003), and when new preferences are being generated, that is, how much did you like TV Show Y, which you just viewed during the experimental session (Adomavicius et al. 2013).

This latter result is important with respect to a common mechanism posited for the seemingly related phenomenon of anchoring-and-adjustment effects in judgment. In decision research, a systematic bias has been observed that judgments tend to be skewed toward an initial anchor value. For example, when Tversky and Kahneman (1974) asked participants to guess the percentage of African countries that were members of the United Nations, those who were asked "Was it more or less than 45%?" guessed lower values than those asked "Was it more or less than 65%?" Most of this research has, like this example, involved participants responding from memory to questions of objective fact (see review by Chapman and Johnson 2002). In contrast, recommender system effects are preference-based where no objective standard is available; and they apply to new preferences as well as recalled preferences. For recall, a common mechanism is *uncertainty*. (How much did I like that movie I saw a year ago?) When asked for a preference, the user starts with the anchor and responds with the first plausible value from a distribution of an uncertain preference. Starting with a low anchor (provided by the recommender system prediction), one tends to arrive at a lower plausible response than when starting at a high anchor. However, this uncertainty mechanism does not operate so convincingly in the case of preference construction. For example, how much uncertainty does one have about a TV show or joke they just experienced? Thus, other mechanisms must be at play either in addition to, or instead of, this mechanism of uncertainty in judgment.

Generalizations

Having established the existence of preference pollution, one set of immediate questions concerns the generalizability of the phenomenon. The robustness of biases, as just discussed, partially speaks to this issue; but other questions of generalizability also arise. Two such questions that have been investigated are discussed in this section. These questions extend the study of the preference pollution phenomenon to a broader scope of settings and interfaces, thereby expanding our understanding of how general the phenomenon is.

System Biases and Economic Behavior. The first question that arises from the previously mentioned studies is: *Could it be that preference pollution is just an artifact of a lack of incentives?* It might be argued that the user ratings that are derived in the experiment have no use outside of the experimental session. They are not like online ratings at a retail site where the ratings provide inputs to a system with which the user has ongoing contact. Thus, the experimental ratings might be seen by the users as providing no discernible extrinsic incentive value.

However, in a different study, online recommendations were found to substantially affect not only users' self-reported preference ratings/opinions about items but also how much consumers were willing to actually pay for them (Adomavicius et al. 2018). In a set of experiments, participants were asked to purchase digital songs that were presented with a series of recommendations. The researchers used modern recommendation algorithms to predict the participants' preferences for individual songs and then either manipulated the recommendations, by perturbing them upward or downward, or left them unmanipulated. Participants were able to purchase the recommended songs by naming their own price. Results showed that perturbed recommendations displayed to participants significantly pulled their willingness to pay in the direction of the recommendation. Based on comparisons of

randomly applied upward and downward perturbations, it was observed that increasing the recommendation rating by one star consistently and systematically increased the willingness to pay by almost 20%.

Thus, preference pollution extends beyond the effects on user ratings to influence economic actions and purchasing behavior where real incentives exist. Again, related to the "uncertain preferences" explanation discussed above, this effect arises even when the users are forced to sample songs prior to making willingness-to-pay judgments (i.e., post-consumption). So, even when the role of uncertainty of song preference is reduced, the observed recommendation effects persist unabated (Adomavicius et al. 2018).

In addition, since the system provided ratings using a 1–5 star rating scale, whereas users responded with song prices on a 0–99 cent scale, another proposed mechanism for anchoring effects is also disconfirmed: a scale-compatibility mechanism. The scale-compatibility explanation (Tversky et al. 1988) argues that, the more compatible the scales of the anchor and the response (e.g., both measured in star-rating points on a 1–5 scale), the higher the weight of the anchor in the decision process. Tversky and Kahneman (1974) hypothesized this explanation for the related anchoring phenomenon. However, as shown in (Adomavicius et al. 2018), even when the scales differ in the recommender system context (1–5 stars for ratings vs. 0–99 cents for price), the effect persists.

System Biases and Non-Personalized Ratings. In addition to *personalized ratings* for products (representing estimated preferences of individual users), online word-of-mouth in the form of *aggregate ratings* for products (representing population-level preference consensus) represents another important type of information on which consumers often rely to make their product decisions. While in many cases the information is presented essentially in the same form (i.e., as numeric ratings), there is an underlying difference in meaning between the mean of aggregated user ratings and personalized system recommendations. In particular, average ratings have a substantial human and social component, and they do not represent personalized information (as recognized, e.g., in Chen et al. 2020 under the characterization of a conformity bias).

According to a study that examined recommender system biases in the context of joke recommendations, users report significantly inflated preference ratings after observing high (as compared to low) rating values, regardless of whether the presented ratings were aggregate user ratings or recommender system predicted, personalized ratings (Adomavicius et al. 2022). Even though the two types of ratings represent very different information, they both tend to generate biases of comparable magnitude when displayed individually. Interestingly, however, when aggregate and personalized ratings are presented together (regardless of their order), they do not generate cumulative (additive) anchoring effects, but exhibit about the same effect magnitude as generated by either aggregate or by personalized ratings alone. Additional experimental evidence suggests that, when both types of ratings are present, personalized ratings seem to be taken into account by users more strongly (and, hence, are more influential in generating biases) than aggregate ratings. The robust result further emphasizes the persistence and dominance of the preference pollution effect. An interesting direction for future work would be to investigate whether other system-provided information (i.e., beyond personalized or aggregate ratings) could potentially lead to preference pollution effects, such as textual information (e.g., reviews), summary statistics (Coba et al. 2020), and social influence factors.

Mitigating system biases

Given that preference pollution exists and that it is a robust phenomenon, a natural question arises as to whether anything can be done about it. The goal is to achieve in practice the preference independence that is widely presumed to exist between observed recommendations and post-consumption preference ratings, or at least to reduce the degree of non-independence. As recommender systems become increasingly popular in today's online environments, preventing or reducing preference pollution constitutes an important research problem. Two broad strategies for addressing this problem are: (i) "modifying the decision environment" (Soll et al. 2016) so that the biases are prevented or reduced proactively, at the ground truth data collection time, or (ii) reducing biases in data after the fact, that is, computationally.

A recent study tried the approach of proactively preventing the biasing effects of recommender systems via user-interface-based solutions that attempt to reduce the biases at rating collection time (Adomavicius et al. 2019). Seven different rating display designs for communicating recommendations to the user were tested, all connected to designs or design aspects that are used in practice. Importantly, none of the seven rating display options completely removed the preference pollution biases generated by recommendations. However, some interface displays were more advantageous than others for reducing the effects. In particular, *graphical* recommendation displays led to significantly lower biases than equivalent numeric forms when users were responding with the typical numerical 1–5 star preference rating scale. Two separate mechanisms were hypothesized to be driving this effect: scale compatibility and differential processing/absorption of graphical versus numeric information. The study found consistent

evidence for a scale compatibility mechanism at work, such that bias is greater when the recommendation display format is the same as the user response format. Only partial (i.e., less robust) evidence was found for the alternative mechanism (differential processing being elicited by graphical as opposed to numerical information displays) as also operating. Overall, it has been shown that scale compatibility is not necessary for preference pollution to occur but, when present, it contributes to preference pollution. This strongly suggests that preference pollution may have multiple precipitants, which increases the mitigation challenge.

Thus, the display of system-predicted preference ratings (in multiple formats) as item recommendations has been shown to bias users' post-consumption preference ratings in the direction of the predicted rating. Top-N lists represent another common approach for presenting item recommendations in recommender systems, as investigated by Adomavicius et al. (2021). The measure of preference pollution with top-N lists is different since the usual comparison of receiving random high and low recommendations is not applicable. Using the bottom-N lists as a comparison is not a plausibly realistic option. Instead, the methodology involved comparisons of lists identified as top-N lists with lists that are not so identified. To bolster the comparison to previous studies, the researchers also looked at lists where the items in the list were shown with system predicted ratings, as compared to lists that did not show these predicted ratings for the items. It turns out that top-N lists with explicit ratings shown for the items replicated the usual effect of preference pollution. However, top-N lists *without explicit rating information* do not induce a discernible bias in subsequent user preference judgments. This result is robust, holding for both lists of personalized item recommendations and non-personalized lists of items based on aggregate user ratings. This suggests that the biasing effect arises primarily or exclusively from the explicit rating information. Simply identifying items as being relevant to a user (e.g., as part of the top-N list) is not enough to induce preference pollution.

There have also been attempts to use explicit debiasing strategies on historical (i.e., already collected) rating data. One example of such work is (Zhang et al. 2017), which analyzes the potential impact of "user expectations" and "item quality" (both of which are empirically estimated from rating data) on the rating of the currently consumed product. In particular, inspired by several psychological theories, the authors conjecture the following assimilation-contrast effects as part of user behavior: users either "assimilate" (conform) to historical ratings (which represent user expectations) if these ratings are not far from the item quality, or users "contrast" (deviate) from historical ratings if these ratings are significantly different from the item quality. Accordingly, the authors propose an algorithmic approach that empirically demonstrates some performance improvements in making more accurate rating predictions. An interesting study for future work would be to validate the conjectured "assimilation" and "contrast" mechanisms using lab or field experiments.

Summary

Table 1 presents a high-level summary of what is known about the preference pollution effects of system recommendations on user preference responses, both in terms of user ratings and economic judgments. The effects are robust across multiple studies involving a variety of digital goods, and for artificial and perturbed system predictions. The effects are observed even immediately after consumption of the item being rated by the user, where preference uncertainty is minimal, at best. The robustness to varying conditions also informs the non-necessity of certain different mechanisms for the effect, for example, preference uncertainty and scale compatibility. Robustness is also demonstrated by the dominance of personalized system biases in the presence of the non-personalized rating information as well as by the difficulty in removing the effects by changing the interface design, though the effects can be somewhat mitigated.

CONCLUSIONS AND FUTURE DIRECTIONS

Having high-quality "ground truth" data is crucial for the predictive models underlying recommender systems. Unfortunately, in addition to a number of biases that online recommendations are known to manifest (Baeza-Yates 2020; Chen et al. 2020), another inherent feature of recommender systems and the iterative user-recommender interactions is that they pose significant challenges to both *preference representativeness* and *preference independence* characteristics of ground truth data. The former is not surprising, as recommender systems are explicitly designed to affect users' consumption choices. Failures of representativeness have been well understood and extensively studied in recommender systems literature and beyond, with a number of mitigation strategies proposed. However, the failures of preference independence represent an unintended preference pollution side effect, likely due to human behavioral decision-making biases, which has been relatively underexplored in recommender systems literature. Thus, this paper places an emphasis on the preference pollution aspect in its overview of the fundamental issue of how recommender systems inherently affect ground truth preference data.

TABLE 1 Summary of preference pollution effects

Preference pollution: What is it? Preference pollution occurs when the preference rating a user provides after consuming an item is systematically impacted by the system recommendation observed pre-consumption. Predictive model-based recommender systems typically rely on the assumption that the users' self-reported preference ratings represent ground truth and are independent of the system recommendations presented.

Preference pollution: Does it occur? YES

Does it generalize? YES

It is a consistent and robust finding that occurs in a variety of settings, where system-predicted preference ratings are displayed as part of the recommendation:

- Types of items (TV Shows, Music, Movies, Jokes)
- Nature of manipulation
  - Artificial recommender system ratings
  - Perturbed recommender system ratings
- Timing of the user's rating task
  - Preference recall (e.g., did you like that movie you saw a year ago?)
  - Preference construction (e.g., did you like the movie you just finished watching?)
- Type of user's preference response
  - Preference ratings (judged relevance of items)
  - Willingness-to-pay judgments (economic behavior)
- Type of recommendation
  - Personalized system-predicted ratings
  - Non-personalized, aggregate (average) user ratings

Posited explanations/mechanisms

- Uncertainty (preference is uncertain, and user provides a rating by adjusting from the recommendation to the first plausible value)
- Biased recall (the recommendation leads to recall of elements of a past experience that are consistent with the recommendation)
- Priming (seeing the recommendation prior to consumption predisposes the user's attitude toward the subsequent experience)
- Integration of relevant/trustworthy information (to the extent users trust the recommendation, they tend to use it as informative of a "correct" preference)
- Scale compatibility (the effect results from an adjustment process whereby the more compatible the scales of the recommendation and the response, the higher the weight of the effect of the recommendation)

Can It be mitigated? Partially/potentially

- Proactive preference pollution has not been eliminated in any setup where the predicted rating information is displayed to the user
- Initial studies show that it can be somewhat reduced by:
  - Using graphical recommendation/rating displays
  - Breaking the scale compatibility in the format of the system's recommendation versus user response
- No preference pollution is observed with top-N recommendation lists *without* explicit rating information
- Some heuristic computational debiasing strategies for reducing preference pollution *after the fact* (in historical rating data) show promise

As recommender systems become ubiquitous, the presence and prevalence of the preference pollution phenomenon has several important implications. First, as online systems make recommendations based on users' feedback, the bias introduced in users' post-consumption preference ratings contaminates the data used by the recommender system, thus, potentially reducing its effectiveness for future recommendations. Second, this provides opportunities for various kinds of unscrupulous and manipulative behavior. Therefore, users may need to become more cognizant of the potential decision-making biases introduced through online recommendations. Just as savvy consumers understand the impacts of advertising, discounting, and pricing strategies, they may also need to consider the potential impact of recommendations on their selection, purchasing, and consumption decisions. In summary, the design and application of next-generation recommendation systems would benefit significantly from considering the impact of preference pollution effects on user behavior.

As to the future, many questions remain unanswered, such as whether users can psychologically distinguish between the pre-consumption value of recommendations and the post-consumption preference task. In other words, if we reduce the bias created by system recommendations upon preference ratings after-the-fact, we do not want to reduce the usefulness of the recommendations before-the-fact. Conceptually, the two are separable; however, it is an open question as to whether they are separable psychologically. Is preference pollution an unavoidable consequence of providing useful recommendations? This is an important question for future studies, particularly given there

has been some success at reducing the biases created by system recommendations. In practice, new interface designs, such as Netflix's move to a "percent match" recommendation and a "thumbs up/down" preference rating, may create interesting further avenues for testing strategies for mitigating these preference pollution biases.

Also, while most prior studies largely focused on observing the immediate/short-term effects on user preferences and behavior, little research has considered the *persistence* of preference pollution. Understanding whether and to what extent biases persist after experiencing a delay (e.g., 1 day, 3 days, 1 week, 1 month) following their exposure to a recommendation would provide important theoretical insights and practical implications.

Similarly, since the user-recommender interactions are dynamic and iterative, it is worthwhile to explore how the preference biasing effects evolve over time and dynamically influence the recommender systems' performance, users' preference ratings, and item consumptions. For example, as users provide more feedback ratings, do the biases accumulate in the system and lead to increasingly worse recommendations and larger preference biases? Or is the system capable of self-correcting over time? Future research can usefully explore the longitudinal impacts of preference bias on recommender systems (e.g., predictive accuracy, recommendation diversity), users (e.g., consumption choices, reliance on the system, trust in the system), and items (e.g., consumption distribution).

Most importantly, because predictive modeling of user preferences (using various statistical and machine learning techniques) is at the heart of modern recommender systems, better and more nuanced understanding of these inherent preference biases and their mitigation should have significant impact on the design of next-generation recommendation techniques and their performance.

CONFLICT OF INTEREST

The authors declare that there is no conflict.

ORCID

Gediminas Adomavicius https://orcid.org/0000-0001-5251-5098
Jesse C. Bockstedt https://orcid.org/0000-0002-4274-9744
Shawn P. Curley https://orcid.org/0000-0002-1982-5534
Jingjing Zhang https://orcid.org/0000-0002-6805-8685

REFERENCES

Abdollahpouri, H. 2019. "Popularity Bias in Ranking and Recommendation." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Honolulu, HI, USA: Association for Computing Machinery, 529–530.

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., and Pizzato, L. 2020a. "Multi-stakeholder Recommendation: Survey and Research Directions." *User Modeling and User-Adapted Interaction* 30(1): 127–158.

Abdollahpouri, H., Burke, R., and Mobasher, B. 2017. "Controlling Popularity Bias in Learning-to-Rank Recommendation." In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 42–46.

Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. 2020b. "The Connection between Popularity Bias, Calibration, and Fairness in Recommendation." In *Fourteenth ACM Conference on Recommender Systems*. Virtual Event. Brazil: Association for Computing Machinery, 726–731.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2013. "Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects." *Information Systems Research* 24(4): 956–975.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2018. "Effects of Online Recommendations on Consumers' Willingness to Pay." *Information Systems Research* 29(1): 84–102.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2019. "Reducing Recommender Systems Biases: An Investigation of Rating Display Designs." *Management Information Systems Quarterly* 43(4): 1321–1341.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2021. "Effects of Personalized and Aggregate Top-N Recommendation Lists on User Preference Ratings." *ACM Transactions on Information Systems (TOIS)* 39(2): 1–38.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2022. "Effects of Personalized Versus Aggregate Ratings on Consumer Preference Responses." *Management Information Systems Quarterly* 46(1): 627–643.

Adomavicius, G., and Kwon, Y. 2012. "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques." *IEEE Transactions on Knowledge and Data Engineering* 24(5): 896–911.

Ariely, D. 2010. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. HarperCollins Publishers.

Azaria, A., Hassidim, A., Kraus, S., Eshkol, A., Weintraub, O., and Netanely, I. 2013. "Movie Recommender System for Profit Maximization." In *Proceedings of the 7th ACM conference on Recommender systems*. Hong Kong, China: ACM, pp. 121–128.

Baeza-Yates, R. 2018. "Bias on the Web." *Communications of the ACM* 61(6): 54–61.

Baeza-Yates, R. 2020. "Bias in Search and Recommender Systems." In *Fourteenth ACM Conference on Recommender Systems*. Virtual Event. Brazil: Association for Computing Machinery, p. 2.

Bertsimas, D., Pawlowski, C., and Zhuo, Y.D. 2017. "From Predictive Methods to Missing Data Imputation: An Optimization Approach." *Journal of Machine Learning Research* 18(1): 7133–7171.

Burke, R., Sonboli, N., and Ordonez-Gauger, A. 2018. "Balanced Neighborhoods for Multi-Sided Fairness in Recommendation." In *Conference on Fairness, Accountability and Transparency*: PMLR, pp. 202–214.

Chapman, G., and Johnson, E. 2002. "Incorporating the Irrelevant: Anchors in Judgments of Belief and Value." In *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich, D. Griffin and D. Kahneman (eds.). Cambridge: Cambridge University Press, pp. 120–138.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. 2020. "Bias and Debias in Recommender System: A Survey and Future Directions." arXiv preprint *arXiv:2010.03240*.

Coba, L., Rook, L., and Zanker, M. 2020. "Choosing between Hotels: Impact of Bimodal Rating Summary Statistics and Maximizing Behavioral Tendency." *Information Technology & Tourism* 22(1): 167–186.

Collins, A., Tkaczyk, D., Aizawa, A., and Beel, J. 2018. "Position Bias in Recommender Systems for Digital Libraries." In *International Conference on Information*. Springer International Publishing AG, pp. 335–344. https://doi.org/10.1007/978-3-319-78105-1_37

Cosley, D., Lam, S., Albert, I., Konstan, J.A., and Riedl, J. 2003. "Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions." In *Conference on Human Factors in Computing Systems*. Fort Lauderdale, FL: ACM New York, NY, pp. 585–592.

Ekstrand, M.D., Burke, R., and Diaz, F. 2019. "Fairness and Discrimination in Recommendation and Retrieval." In *Proceedings of the 13th ACM Conference on Recommender Systems*. September 16–20, 2019, Copenhagen, Denmark, pp. 576–577.

Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., and Pera, M.S. 2018. "All the Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness." In Conference on Fairness, Accountability and Transparency: PMLR, pp. 172–186.

Flaxman, S., Goel, S., and Rao, J.M. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80(S1): 298–320.

Gao, R., and Shah, C. 2020. "Counteracting Bias and Increasing Fairness in Search and Recommender Systems." In *Fourteenth ACM Conference on Recommender Systems*. Virtual Event. Brazil: Association for Computing Machinery, pp. 745–747.

Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., and Zhang, Y. 2020. "Understanding Echo Chambers in E-Commerce Recommender Systems." In *Proceedings of the 43rd International Acm Sigir Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, pp. 2261–2270.

Guo, H., Yu, J., Liu, Q., Tang, R., and Zhang, Y. 2019. "Pal: A Position-Bias Aware Learning Framework for Ctr Prediction in Live Recommender Systems." In *Proceedings of the 13th ACM Conference on Recommender Systems*. Copenhagen, Denmark: Association for Computing Machinery, pp. 452–456.

Hofmann, K., Schuth, A., Bellogin, A., and De Rijke, M. 2014. "Effects of Position Bias on Click-Based Recommender Evaluation." In European Conference on Information Retrieval. Springer, pp. 624–630.

Ishioka, T. 2014. "Investigations into Missing Values Imputation Using Random Forests for Semi-Supervised Data." In *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services*. Hanoi, Viet Nam: Association for Computing Machinery, pp. 296–301.

Jannach, D., and Adomavicius, G. 2017. "Price and Profit Awareness in Recommender Systems." arXiv preprint *arXiv:1707.08029*.

Jannach, D., and Bauer, C. 2020. "Escaping the Mcnamara Fallacy: Towards More Impactful Recommender Systems Research." *AI Magazine* 41(4): 79–95.

Jannach, D., Karakaya, Z., and Gedikli, F. 2012. "Accuracy Improvements for Multi-Criteria Recommender Systems." In *Proceedings of the 13th ACM Conference on Electronic Commerce*. Valencia, Spain: ACM, pp. 674–689.

Jannach, D., Lerche, L., Kamehkhosh, I., and Jugovac, M. 2015. "What Recommenders Recommend: An Analysis of Recommendation Biases and Possible Countermeasures." *User Modeling and User-Adapted Interaction* 25(5): 427–491.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kim, Y.-D., and Choi, S. 2014. "Bayesian Binomial Mixture Model for Collaborative Prediction with Non-Random Missing Data." In *Proceedings of the 8th ACM Conference on Recommender Systems*. Foster City, Silicon Valley, California, USA: ACM, pp. 201–208.

Lakiotaki, K., Matsatsinis, N.F., and Tsoukiàs, A. 2011. "Multicriteria User Modeling in Recommender Systems." *IEEE Intelligent Systems* 26(2): 64–76.

Lim, D., McAuley, J., and Lanckriet, G. 2015. "Top-N Recommendation with Missing Implicit Feedback." In *Proceedings of the 9th ACM Conference on Recommender Systems*. Vienna, Austria: ACM, pp. 309–312.

Little, R.J., and Rubin, D.B. 2019. *Statistical Analysis with Missing Data*. John Wiley & Sons.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., and Burke, R. 2020. "Feedback Loop and Bias Amplification in Recommender Systems." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, pp. 2145–2148.

Marlin, B., Zemel, R.S., Roweis, S., and Slaney, M. 2012. "Collaborative Filtering and the Missing at Random Assumption." arXiv preprint *arXiv:1206.5267*.

Marlin, B.M., and Zemel, R.S. 2009. "Collaborative Prediction and Ranking with Non-Random Missing Data." In *Proceedings of the Third ACM Conference on Recommender Systems*. New York, USA: Association for Computing Machinery, pp. 5–12.

Mobasher, B., Burke, R., Bhaumik, R., and Williams, C. 2007. "Toward Trustworthy Recommender Systems: An Analysis of Attack Models and Algorithm Robustness." *ACM Transactions on Internet Technology* 7(4): 23:21—23:38.

Nguyen, T.T., Hui, P.-M., Harper, F.M., Terveen, L., and Konstan, J.A. 2014. "Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity." In *Proceedings of the 23rd International Conference on World Wide Web*. Seoul, Korea: ACM, pp. 677–686.

Nugroho, H., and Surendro, K. 2019. "Missing Data Problem in Predictive Analytics." In *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. Penang, Malaysia: Association for Computing Machinery, pp. 95–100.

Pariser, E. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.

Pradel, B., Usunier, N., and Gallinari, P. 2012. "Ranking with Non-Random Missing Ratings: Influence of Popularity and Positivity on Evaluation Metrics." In *Proceedings of the sixth ACM conference on Recommender systems*. Dublin, Ireland: ACM, pp. 147–154.

Prawesh, S., and Padmanabhan, B. 2014. "The "Most Popular News" Recommender: Count Amplification and Manipulation Resistance." *Information Systems Research* 25(3): 569–589.

Saito, Y. 2020. "Asymmetric Tri-Training for Debiasing Missing-Not-at-Random Explicit Feedback." In *Proceedings of the 43rd International Acm Sigir Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, pp. 309–318.

Saito, Y., Yaginuma, S., Nishino, Y., Sakata, H., and Nakata, K. 2020. "Unbiased Recommender Learning from Missing-Not-at-Random Implicit Feedback." In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. Houston, TX, USA: Association for Computing Machinery, pp. 501–509.

Santore, F., Almeida, E.C.d., Bonat, W.H., Pena, E.H.M., and Oliveira, L.E.S.d. 2020. "A Framework for Analyzing the Impact of Missing Data in Predictive Models." In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, pp. 2209–2212.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. 2016. "Recommendations as Treatments: Debiasing Learning and Evaluation." arXiv preprint *arXiv:1602.05352*.

Soll, J.B., Milkman, K.L., and Payne, J.W. 2016. "A User's Guide to Debiasing." In *The Wiley Blackwell Handbook of Judgment and Decision Making*. John Wiley & Sons, Ltd, pp. 924–951.

Steck, H. 2010. "Training and Testing of Recommender Systems on Data Missing Not at Random." In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, DC, USA: Association for Computing Machinery, pp. 713–722.

Steck, H. 2013. "Evaluation of Recommendations: Rating-Prediction and Ranking." In *Proceedings of the 7th ACM Conference on Recommender Systems*. Hong Kong, China: ACM, pp. 213–220.

Sürer, Ö., Burke, R., and Malthouse, E.C. 2018. "Multistakeholder Recommendation with Provider Constraints." In *Proceedings of the 12th ACM Conference on Recommender Systems*. Vancouver, British Columbia, Canada: Association for Computing Machinery, pp. 54–62.

Tremblay, M.C., Dutta, K., and Vandermeer, D. 2010. "Using Data Mining Techniques to Discover Bias Patterns in Missing Data." *Journal of Data and Information Quality* 2(1): 2.

Tversky, A., and Kahneman, D. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* (185): 1124–1131.

Tversky, A., Sattath, S., and Slovic, P. 1988. "Contingent Weighting in Judgement and Choice." *Psychological Review* 95(3): 371–384.

Wang, X., Golbandi, N., Bendersky, M., Metzler, D., and Najork, M. 2018. "Position Bias Estimation for Unbiased Learning to Rank in Personal Search." In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Marina Del Rey, CA, USA: Association for Computing Machinery, pp. 610–618.

Wang, X., Zhang, R., Sun, Y., and Qi, J. 2019. "Doubly Robust Joint Learning for Recommendation on Data Missing Not at Random." In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California, pp. 6638–6647.

Yang, L., Cui, Y., Xuan, Y., Wang, C., Belongie, S., and Estrin, D. 2018. "Unbiased Offline Recommender Evaluation for Missing-Not-at-Random Implicit Feedback." In *Proceedings of the 12th ACM Conference on Recommender Systems*. October 2–7, 2018, Vancouver, BC, Canada, pp. 279–287.

Zanker, M., Rook, L., and Jannach, D. 2019. "Measuring the Impact of Online Personalisation: Past, Present and Future." *International Journal of Human-Computer Studies* (131): 160–168.

Zhang, J., Adomavicius, G., Gupta, A., and Ketter, W. 2020. "Consumption and Performance: Understanding Longitudinal Dynamics of Recommender Systems Via an Agent-Based Simulation Framework." *Information Systems Research* 31(1): 76–101.

Zhang, X., Zhao, J., and Lui, J.C. 2017. "Modeling the Assimilation-Contrast Effects in Online Product Rating Systems: Debiasing and Recommendations." In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. August 27–31, 2017, Como, Italy, pp. 98–106.

Zheng, Y. 2019. "Multi-Stakeholder Recommendations: Case Studies, Methods and Challenges." In *Proceedings of the 13th ACM Conference on Recommender Systems*. Copenhagen, Denmark: Association for Computing Machinery, pp. 578–579.

Zheng, Y., Ghane, N., and Sabouri, M. 2019. "Personalized Educational Learning with Multi-Stakeholder Optimizations." In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. Larnaca, Cyprus: Association for Computing Machinery, pp. 283–289.

AUTHOR BIOGRAPHIES

Gediminas Adomavicius is a professor in the Department of Information and Decision Sciences at the University of Minnesota. He received his PhD degree in computer science from New York University. His general research interests are in computational techniques for aiding decision-making in information-intensive environments and include personalization technologies and recommender systems, machine learning and data analytics, and electronic market mechanisms. He has published in a number of leading academic journals in information systems and computer science, has received several major research grants (including the *U.S. National Science Foundation* CAREER award), has served on the editorial boards of several journals (including as Senior Editor of *Information Systems Research* and *MIS Quarterly*), and is a Distinguished Fellow of INFORMS Information Systems Society.

Jesse Bockstedt is a professor of information systems and operations management at the Goizueta Business School of Emory University. He received his PhD from the Carlson School of Management at the University of Minnesota. He studies user behavior and economic issues in environments that rely on information technology. His research has appeared in leading academic journals including *Information Systems Research*, *MIS Quarterly*, *Journal of MIS*, *Journal of Operations Management*, and *Production and Operations Management*.

Shawn Curley is a Professor of Information and Decision Sciences at the Carlson School of Management, University of Minnesota – Twin Cities. He received his PhD in Psychology from the University of Michigan, Ann Arbor in 1986. His research interests include decision and judgment processes under uncertainty, the use of personalization technology, behavior in combinatorial multi-item auctions, measures of uncertainty, and medical decision making.

Jingjing Zhang is an associate professor of Information Systems and a Fettig/Whirlpool Faculty Fellow at the Kelley School of Business, Indiana University. She received her PhD from the Carlson School

of Management at the University of Minnesota. Her research interests include personalization techniques, recommender systems, and human-computer interactions. Her research has appeared in leading academic journals including *MIS Quarterly*, *Information Systems Research*, *INFORMS Journal on Computing*, *IEEE Transactions on Knowledge and Data Engineering*, and *ACM Transactions on Information Systems*.

How to cite this article: Adomavicius, G., Bockstedt, J.C., Curley, S.P., and Zhang, J. 2022. "Recommender systems, ground truth, and preference pollution." *AI Magazine* 43: 177–89. https://doi.org/10.1002/aaai.12055