


The Editor and the Algorithm: Recommendation Technology in Online News

Journal Article

Author(s):

Peukert, Christian ; Sen, Ananya; Claussen, Jörg

Publication date:

2024-09

Permanent link:

<https://doi.org/10.3929/ethz-b-000655706>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

Management Science 70(9), <https://doi.org/10.1287/mnsc.2023.4954>

The Editor and the Algorithm: Recommendation Technology in Online News

Christian Peukert,^a Ananya Sen,^{b,*} Jörg Claussen^c

^aFaculty of Business and Economics, University of Lausanne, 1015 Lausanne, Switzerland; ^bCarnegie Mellon University Pittsburgh, Pennsylvania 15213; ^cLMU Munich, 80539 München, Germany and Copenhagen Business School, 2000 Frederiksberg, Denmark

*Corresponding author

Contact: christian.peukert@unil.ch,  <https://orcid.org/0000-0003-3997-8850> (CP); ananyase@andrew.cmu.edu,  <https://orcid.org/0000-0002-9082-6871> (AS); j.claussen@lmu.de,  <https://orcid.org/0000-0001-8432-8860> (JC)

Received: November 21, 2019

Revised: January 6, 2021; May 18, 2022;
December 9, 2022; January 20, 2023

Accepted: February 8, 2023

Published Online in Articles in Advance:
October 17, 2023

<https://doi.org/10.1287/mnsc.2023.4954>

Copyright: © 2023 The Author(s)

Abstract. We run a field experiment to study the relative performance of human curation and automated personalized recommendation technology in the context of online news. We build a simple theoretical model that captures the relative efficacy of personalized algorithmic recommendations and curation based on human expertise. We highlight a critical tension between detailed, yet potentially narrow, information available to the algorithm versus broad (often private), but not scalable, information available to the human editor. Empirically, we show that, on average, algorithmic recommendations can outperform human curation with respect to clicks, but there is significant heterogeneity in this treatment effect. The human editor performs relatively better in the absence of sufficient personal data and when there is greater variation in preferences. These results suggest that reverting to human curation can mitigate the drawbacks of personalized algorithmic recommendations. Our computations show that the optimal combination of human curation and automated recommendation technology can lead to an increase of up to 13% in clicks. In absolute terms, we provide thresholds for when the estimated gains are larger than our estimate of implementation costs.

History: Accepted by Chris Forman, information systems.



Open Access Statement: This work is licensed under a Creative Commons Attribution 4.0 International License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “Management Science. Copyright © 2023 The Author(s). <https://doi.org/10.1287/mnsc.2023.4954>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by/4.0/>.”

Funding: C. Peukert acknowledges funding from the Swiss National Science Foundation [Grant No. 100013_197807].

Supplemental Material: The e-companion and data files are available at <https://doi.org/10.1287/mnsc.2023.4954>.

Keywords: online news • human expertise • technology adoption • algorithmic recommendations • data

1. Introduction

Recommender systems are prominent machine learning (ML) applications in online platforms, as they promise to automate tasks historically carried out by humans. There is significant discourse around the potential of personalized algorithmic recommendation to shore up online revenues vis-à-vis human decisions, particularly in the news industry (Bodó 2019, Gulla et al. 2021). Recommender systems can take over the tasks of editors who select the news stories shown to readers (Gulla et al. 2021). A key tension arises when comparing automation to human curation. Algorithmic recommendations personalize at scale using information that tends to be detailed but (often) temporally narrow and context specific. On the other hand, human experts base recommendations on broad knowledge

accumulated over the course of a professional career; hence, they can have comparative benefits because of private information but cannot make individual recommendations at scale. However, there is little empirical evidence on how the tension between scalable detailed (yet narrow) and broad (private) unscalable information plays out. Additionally, in contexts other than the news industry, there is some application of human curation alongside automated recommendations on online platforms. Amazon, for example, offers automated recommendations (“People who bought this also bought”) as well as human-curated lists of products (“Editors’ Picks”). YouTube mainly uses algorithmic recommendations to steer users to content, but users can also create playlists and curate content for other users. Similarly, Spotify’s algorithm creates fully

automated personalized recommendations (“Discover Weekly”) but also hosts a large variety of curated playlists (Aguilar and Waldfogel 2021). Despite the managerial importance of understanding the economic value of recommender systems relative to human curation to optimize technology adoption decisions, limited systematic evidence exists.

We study how users react to automated recommendations relative to human curation in a large-scale field experiment we conduct with a major news outlet in Germany. Human editors select the order of articles that appear on the outlet’s homepage. Generally, any user who arrives at the homepage sees the same content in the same place. In the experiment, a user is randomly assigned to the human-curated version or a treatment condition every time the user visits the homepage. In the treatment condition, we customize the homepage to show automated recommendations from an algorithm trained on data capturing the browsing behavior of that individual user.

To guide the empirical analysis, we develop a simple theoretical model that contrasts humans with recommendation algorithms in their ability to predict consumer preferences. The model shows that differences in the types and volume of information a human editor and a recommendation algorithm can access explain relative prediction quality. We characterize situations where automated recommendation can outperform human curation, depending on the amount of accessible individual data and variation in consumer preferences. Additionally, using this framework, we derive testable predictions about the relative efficacy based on the stock versus the flow of personal data and the potential for the optimal strategy to be a combination of the human editor and the algorithm.

The empirical results validate the predictions of the theoretical model. First, we show that, on average, the algorithm outperforms the human, but there is a significant amount of heterogeneity with respect to the amount of user-level data available. The analysis demonstrates that the algorithm requires a sufficient volume of engagement data, measured by the prior number of visits to the website, from an individual to outperform the human editor. In the absence of personal data, the algorithm in our field experiment relies on aggregate demand trends to make recommendations and solve its cold-start problem. Human curation outperforms automated recommendations without personal data. Moreover, we show that the value of individual-level information for the algorithm increases at a diminishing rate. Additionally, we provide suggestive evidence that a more experienced human editor (relative to a less experienced human editor) outperforms the algorithm by a greater margin. This finding suggests it is beneficial for a platform to defer to human expertise in the absence of user-specific personal data

despite mitigating the cold-start problem with aggregate data for the algorithm. We show it is optimal to only refer to the algorithm after a user has visited the website a sufficient number of times.

Next, we analyze the dynamics of the relative prediction quality of the algorithm and the human expert based on the stock versus the flow of user-level information available to the algorithm. We leverage a coding bug during the course of the field experiment when the algorithm did not have access to updated personal data for one week. We demonstrate that it is not only the stock but also a flow of updated user-level information that has significant economic implications for the choice of recommendation technology. Clicks in the algorithmic condition become significantly negative relative to human curation during the coding bug period. This negative effect is relatively muted for users who have a higher volume of prior engagement, in other words, a higher stock of individual-level information available to the algorithm. Practically, such system glitches are commonplace (Mehta and Singh 2022), making the strategic decision to revert to human expertise critical for a platform to maintain engagement.

We then turn to our results on the variation in user preferences (because of temporal changes in content availability) and how it affects prediction quality. Human curation performs relatively better on days with more attention-grabbing news, leading to a particularly high variance in clicks per category. Hence, significant variation in the type of content available allows humans to use their domain expertise and adapt to outperform the algorithm. The algorithm, often trained on detailed yet narrow data, cannot adapt when the underlying ground truth changes, and there is limited information on different types of content.

Finally, we perform (counterfactual) computations to estimate the performance of a recommendation system that combines human curation and automated recommendations, taking the strengths and weaknesses of both approaches into account. We find that such a combination would increase clicks for the website by up to 13%. Thus, our results suggest that managers should leverage humans and automated recommendations in tandem rather than looking at the curation problem as human experts versus algorithms.

2. Related Literature

Several papers are directly related to our work (see Online Appendix Table A.1 for a summary). The context ranges from movies (Krishnan et al. 2008, Amatriain et al. 2009, Sinha and Swearingen 2011), jokes (Yeomans et al. 2019), and hiring (Cowgill 2018) to e-commerce (Senecal and Nantel 2004). Each of these papers finds that the optimal recommendation strategy is to choose either the human or the algorithm. We

believe that such implications are a function of the small sample sizes in the studies (Senecal and Nantel 2004, Krishnan et al. 2008, Amatriain et al. 2009, Sinha and Swearingen 2011) and/or the one-shot nature of the experimental settings (Cowgill 2018, Yeomans et al. 2019). Our paper adds nuance by demonstrating, theoretically and empirically, that the optimal strategy in (online) recommendation varies by the amount of available data and variation in user preferences over time. Hence, a combination of humans and algorithms can yield the best results. This result provides a distinct managerial implication relative to the literature and comes out because of the large-scale and long-term field experiment that allows us to uncover key dynamics. Our gold-standard field experiment (Goldfarb and Tucker 2014) lets us measure revealed preferences (clicks) and causal estimates to examine alternative (counterfactual) scenarios, which takes us one step closer to measuring business value. In contrast, the literature often focuses on unincentivized stated outcomes and rarely provides an explicit cost-benefit analysis of adopting recommendation technology.

Further, we relate to papers that analyze how *heterogeneity in human expertise or experience* makes machine recommendations more or less useful. For example, vintage-specific human capital can lead to better use of machine predictions when evaluating patent applications for novelty (Choudhury et al. 2020), and younger players are more proficient in using an AI-powered decision support tool to improve their moves in the game Go (Choi et al. 2022). As demonstrated by chess players (Miric et al. 2020), greater human experience helps to make better decisions when benchmarked against a machine, especially in complex situations. Our paper differentiates from these studies by focusing on variation in critical dimensions of personal data as inputs to ML models while keeping constant the humans and algorithm involved. Indeed, our empirical results viewed through the theoretical framework can shed light on (other) contexts in which humans might have a relative advantage over algorithmic recommendations.

The focus of the related technical literature has been to tweak and compare the relative performance of algorithms even when explicitly analyzing the choice of news stories (Wang et al. 2017). A key differentiator is that we focus on the performance of automated recommendations relative to the status quo of human curation. Moreover, our large-scale field experiment provides *causal* estimates by going beyond offline evaluations, which often form the basis for a majority of applied technical research (e.g., the studies reviewed in Jannach and Jugovac 2019 and Wu et al. 2022). Such offline evaluations often suffer from endogeneity concerns because of the inability of historical data to account for fast-paced dynamics and customer feedback loops (Valavi et al.

2022). More recently, this literature has run into severe methodological issues and problems of replication (Kapoor and Narayanan 2023), highlighting the need for credible *field experiments*.

Our results also relate to policy-related work focusing on algorithms. Studies in this area have aimed to simulate privacy policies where algorithms would not have access to the personal data of users who visit an (online e-commerce) platform (Sun et al. 2023), whereas others have analyzed a competition policy question related to data retention policies using a natural experiment that reduced the time window during which search engines retain individual user data (Chiou and Tucker 2017). Another focus has been to analyze the potential for data network effects in online search as a function of the quantity of past personalized information (Schaefer and Sapi 2020). Unlike these studies, we analyze the complementary role that a human could play when a company adopts algorithmic recommendation technology.

3. Theoretical Model

We now sketch out a simple model of recommendation as a demand prediction exercise, building on frameworks introduced in Peukert and Reimers (2018) and Petrova et al. (2021). Consider a setting with two possible recommendation strategies: human or algorithmic.

The algorithm forms a belief about the utility of an individual consumer $u_i \geq 0$, which is drawn from a normal distribution $u_i \sim N(\mu_i, \sigma_i^2)$. The initial belief comes from granular historical data on the behavior of consumer i , or, if not available, from historical data on the behavior of other consumers, both collected over a relatively short period of time (e.g., months' worth of training data). Because humans cannot process such granular individual-level information at scale, they form a belief that the population average $u > 0$ is drawn from $u \sim N(\mu, \sigma^2)$. The human's belief is based on their subjective observations of aggregate consumer demand over a relatively long time period (e.g., years of professional experience). This setup captures the fact that the algorithm learns from granular, but narrow, information. On the other hand, the human cannot process granular individual-level data but has a broad source of information based on accumulated knowledge. If consumers have heterogeneous preferences, it is possible that the mean utility of any consumer is different from the aggregate mean utility across consumers (i.e., $\mu \neq \mu_i$). For the same reason, the precision of the belief about the individual's utility, defined as $\tau_i = 1/\sigma_i^2 > 0$, can be different from the precision of the belief about the population average utility, $\tau = 1/\sigma^2 > 0$.

The algorithm can learn about consumer i 's utility when consumer i clicks on content. Hence, the algorithm can incorporate individual-level data in future predictions

of u_i in a Bayesian updating process. Let there be $n_i \geq 0$ data points $X_{k,i} = x_{k,i}$ ($k = 1, \dots, n_i$) about consumer i 's behavior, capturing the amount of individual-level information available to the algorithm. Denote the precision of the true distribution of u_i as r_i , which denotes the variation in consumption preferences. The algorithm's posterior is also normally distributed with mean μ'_i and precision $\tau_i + n_i r_i$ (DeGroot 1970, p. 167). In particular, we can write

$$\mu'_i = \frac{n_i r_i}{\tau_i + n_i r_i} x_i + \frac{\tau_i}{\tau_i + n_i r_i} \mu_i, \quad (1)$$

where x_i is the sample mean of the additional data. We assume the sample mean of additional data are greater than zero, implying it has meaningful information.

In contrast to the algorithm, the human can only process aggregate data. To keep the model simple, we assume that the human does not update their prior (does not learn) as aggregate data comes in. This assumption may be equivalent to a human editor trusting their professional experience and intuition rather than monitoring a dashboard with live traffic data. This assumption is not critical for the results but simplifies the computations significantly. A more general version of the model described in the appendix yields qualitatively similar insights. In the generalized version, the human can update their prior with broader, potentially private, information, and we explicitly model the human's learning function based on aggregate information. Overall, the simplified version of the model, as well as the generalized setup in the appendix, capture the intuition that the algorithm is relatively narrow but can be detailed and scaled at the individual level. On the other hand, the human can have broader information, potentially private or not available to the algorithm, but cannot scale similarly.

The human and the algorithm will have beliefs about consumer preferences for each content category, rank them, and recommend content from the category with the highest values of μ and μ'_i , respectively. Consumers respond to the recommendations by clicking on the content if their utility from doing so is sufficiently high, that is, if μ and μ'_i are sufficiently close to the individual's true utility, \tilde{u}_i , of consuming content in that particular category. Predictions of utility per category are, therefore, proportional to performance. Therefore, when we want to compare the performance of algorithmic recommendation and human curation, we can look at the difference in predicted utilities. We define the prior difference as $\delta_i = \mu_i - \mu$ and the posterior difference as

$$\delta'_i = \frac{n_i r_i}{\tau_i + n_i r_i} x_i + \frac{\tau_i}{\tau_i + n_i r_i} \delta_i. \quad (2)$$

In panel (a) of Figure 1, we plot three examples of the posterior difference as given in Equation (2) as a function

of the number of additional data points, n . Positive values on the vertical axis indicate that algorithmic recommendation outperforms human curation, and negative values indicate that human curation outperforms algorithmic recommendation. In all examples the true utility is one, approximated by the sample mean of additional data, $x_i = 1$. The solid line shows an example where the prior of the algorithm is smaller than the prior of the human ($\delta_i = -2$), meaning that the human's prior is closer to the true utility. The dotted line shows an example where the prior of the algorithm is larger than the prior of the human and larger than the sample mean ($x_i < \delta_i = 2$). The dashed line shows an example where the prior of the algorithm is larger than the human prior but smaller than the sample mean ($x_i > \delta_i = 0.01$). These examples illustrate our first two hypotheses.

Hypothesis 1. *Algorithmic recommendation can outperform human curation if there is sufficient individual-specific data.*

Intuition. As seen in panel (a) of Figure 1, the posterior difference is positive if the prior difference is positive. Hence, when the algorithm has a head start and outperforms human curation even with zero additional data points, it will keep outperforming human curation as new data points arrive. However, when the human has a head start, only with a sufficient amount of additional data the algorithm can catch up and eventually outperform human curation. In particular, $\delta'_i > 0$ if $n_i > \frac{\tau_i \delta_i}{r_i x_i}$.

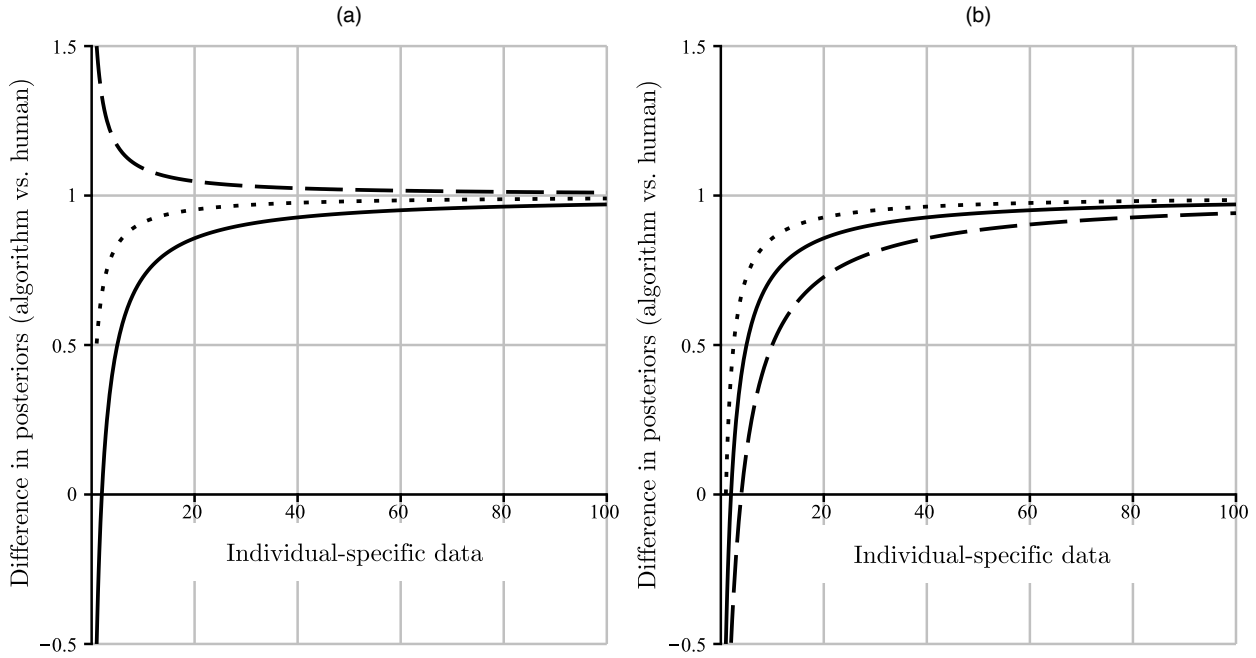
Hypothesis 2. *The relative performance of algorithmic recommendation increases at a diminishing rate with more individual-specific data.*

Intuition. As can be seen in panel (a) of Figure 1, the posterior difference converges toward the true sample mean as more data points arrive. More formally, we can inspect the sign of the first- and second-order derivatives of δ'_i with respect to the number of new data points. In the case that the prior difference is smaller than the sample mean (i.e., neither human nor algorithm overestimates the true utility), more data help to make better predictions, but at a diminishing rate. In particular, $\frac{\partial \delta'_i}{\partial n_i} > 0$ and $\frac{\partial^2 \delta'_i}{\partial n_i^2} < 0$ if $x_i > \delta_i$.

So far, we have only considered the stock of individual-level data, studying comparative statics at different levels of the stock. We now consider explicitly that updating happens over time to analyze the implications of a change in the flow of user data on the relative performance of algorithmic recommendations. The posterior difference at time $t > 0$ can be written as

$$\delta'_{it} = \frac{n_{it} r_i}{\tau_i + n_{it} r_i} x_{it} + \frac{\tau_i}{\tau_i + n_{it} r_i} \delta'_{i,t-1}. \quad (3)$$

For simplicity, assume that in every period, the algorithm can access a fraction $\alpha > 0$ of additional data to

Figure 1. Difference in Algorithm vs. Human as a Function of Individual-Specific Data

Notes. (a) Difference in prior information (δ_i). (b) Variation in consumer preferences (r_i). Both panels plot the difference in algorithmic vs. human posteriors (δ'_i) as a function of the stock of individual-specific data (n_i). We fix values of the precision of the true preference distribution and the sample mean ($\tau_i = 1$ and $x_i = 1$). In (a), we fix a posterior precision value ($r_i = 1$). The solid line in (a) shows an example where the difference in priors is negative ($\delta_i = -2$), the dotted line illustrates a case where the difference in priors is positive and larger than the sample mean ($x_i < \delta_i = 2$), and the dashed line is an example where the difference in priors is positive and smaller than the sample mean ($x_i > \delta_i = 0.01$). In (b), we fix the value of the difference in priors ($\delta_i = -2$) and vary the posterior precision (r_i). The solid line serves as the base case with $r_i = 1$, the dotted line is an example with lower variance ($r_i = 2$), and the dashed line is an example with more variance ($r_i = 0.5$).

update its prior, such that $t_i = \alpha n_{it}$. The fraction α can be interpreted as the flow of data. Then, we can write

$$\delta'_{it} = (\delta_i - x_{it}) \left(\frac{\tau_i}{\tau_i + \alpha n_{it} r_i} \right)^{\alpha n_{it}} + x_{it} \quad (4)$$

Hypothesis 3. *The relative performance of algorithmic recommendation increases faster with a higher frequency of updating, that is, a greater flow of data, but less so with the accumulation of a greater stock of data. That is, there is some degree of substitutability between the stock and the flow of personal data.*

Intuition. With a higher updating frequency, that is, when a greater fraction of the available individual-specific data is used to update a prior in a given period, the posterior approaches the sample mean quicker. To see this, inspect the sign of the derivative of δ'_{it} with respect to α . In particular, $\frac{\partial \delta'_{it}}{\partial \alpha} > 0$ if $\delta_i < 0$ or $x_{it} > \delta_i$ and $\delta_i > 0$. However, over time, a stock of data accumulates. Hence, in each period, a greater fraction of the available data is used to form the posterior, and the rate at which the posterior approaches the sample mean slows down. We can see this by inspecting the cross derivative of δ'_{it} with respect to α and n_{it} . It turns out that $\frac{\partial^2 \delta'_{it}}{\partial \alpha \partial n_{it}} < 0$ again if $\delta_i < 0$ or $x_{it} > \delta_i$ and $\delta_i > 0$ and if, additionally, n_{it} is sufficiently large.

The dimension of stock versus flow is strategically vital because frequently updating data comes with a cost. Hence, data acquisition and management can be a crucial determinant of the economic performance of automated recommendation technology and an important input into the firm technology adoption decision. More broadly, stock versus flow can also affect competition and market structure. When the stock of data is sufficient to predict consumer preferences, an incumbent can have a significant advantage over an entrant. In practice, this might even be relevant with seemingly low stocks of data because it might be very difficult to generate repeated engagement of a sufficiently large number of users. On the other hand, if the flow of data is consequential, then entry barriers are lower.

Having characterized the relative performance of algorithmic recommendation with respect to the stock, flow, and the interaction of stock and flow of individual-level data, we now turn to analyze how the relative performance of algorithmic recommendation varies with respect to the underlying distribution of consumer preferences.

We first turn to the variation of preferences at the individual level, that is, within an individual. In panel (b) of Figure 1, we plot three examples of the posterior difference as given in Equation (2) as a function of the number of additional data points, n . Positive values

on the vertical axis indicate that algorithmic recommendation outperforms human curation; negative values indicate that human curation outperforms algorithmic recommendation. In all examples, the true utility is one, approximated by the sample mean of additional data, $x_i = 1$. We also only look at examples where the prior of the algorithm is smaller than the prior of the human ($\delta_i = -2$), meaning that the human's prior is closer to the true utility. The solid line shows a base case where we set the precision of the underlying distribution of consumer preferences r to one. The dotted line then shows an example with more precision (less variance, $r=2$), and the dashed line shows an example with less precision (more variance, $r=0.5$). We reach positive values on the vertical axis quicker as we reduce variance. That is, when updating the algorithm is more precise, it can outperform human curation with a smaller stock of individual-level data. This leads us to the next hypothesis.

Hypothesis 4. *The relative performance of algorithmic recommendation decreases in an environment characterized by greater variance.*

Intuition. If it is more difficult to precisely predict an individual's preferences with a given data level, the algorithmic recommendation's relative performance is lower. Put differently, any inherent advantage of a human editor in predicting the preferences of the average user is relatively stronger if there is more variation in preferences within an individual consumer. To see this formally, inspect the sign of the derivative of δ'_i with respect to the precision of the posterior distribution. In particular, $\frac{\partial \delta'_i}{\partial r_i} > 0$ if $x_i > \mu_i$.

Finally, we turn to the distribution across consumers to analyze how the relative performance of algorithmic recommendation depends on the distribution of the stock of data. Using the results we have obtained so far, we are interested in characterizing the optimal combination of human curation and algorithmic recommendation depending on users' overall engagement distribution.

Hypothesis 5. *A combination of human curation and algorithmic recommendation leads to higher performance than either alone if consumer engagement is sufficiently heterogeneous.*

Intuition. Consider a policy (θ_i) where human curation is used when it outperforms algorithmic recommendation and vice versa:

$$\theta_i = \begin{cases} \mu & \text{if } \delta'_i \leq 0 \\ \mu'_i & \text{if } \delta'_i > 0, \end{cases} \quad (5)$$

where, as we have shown, $\delta'_i > 0$ if $n_i > \frac{\tau_i \delta_i}{r_i x_i} = \eta_i$. For any share of users, $s \in (0, 1)$, that are sufficiently active on

the website, that is, for which $n_i > \eta_i$, the combination of human curation and algorithmic recommendation yields higher performance than either alone, that is, $\sum_i ((1-s)\mu + s\mu'_i) > \sum_i \mu$ and $\sum_i ((1-s)\mu + s\mu'_i) > \sum_i \mu'_i$. For example, assuming that engagement follows a log-normal distribution with parameters that generate a sufficiently large mass of consumers with relatively low values, the human can outperform the algorithm in the aggregate. In practice, this can be very important if there is a relatively large mass of users that do not show repeated engagement, that is, when customer loyalty is low.

4. Setting and Econometric Model

4.1. Industry Partner

Our industry partner ("NEWS") is one of the largest players in the German news market, with over 20 million monthly visitors and about 120 million monthly page impressions. It is similar to the *Wall Street Journal* in size and influence. Like other major news outlets, NEWS gets the largest share of its revenue from advertising, which makes reader engagement crucial for its financial health. The website does not have a subscription model. Like in other Western democracies, the German news industry has a few major national outlets covering the broad political spectrum. NEWS focuses on politics, finance, and sports while reporting on various other topics.

In general, editorial curation follows an objective function that is hard to characterize explicitly. However, based on the business model of NEWS, it is clear that advertising revenue plays a fundamental role. Anecdotal evidence further suggests that the editorial team monitors analytics dashboards quite closely. The media economics literature highlights that news editors care about increasing advertising revenue explicitly and choose news stories based on those calculations (Gentzkow and Shapiro 2010, Sen and Yildirim 2015 and discussions therein). Moreover, even if editors only care about impact-driven stories or agenda setting (McCombs and Shaw 1972), evidence shows that important stories are highly correlated with greater audience reach as measured by clicks (Sen and Yildirim 2015). Hence, focusing on clicks as the primary outcome variable captures various objectives in a reduced form.

Journalism and data science are in separate divisions at the parent organization of NEWS. The data science team has projects throughout the parent organization (which includes nonjournalistic platforms), whereas journalists work primarily for NEWS. Journalists were not aware of the field experiment. In general, the head of the newsroom leads a team of about six editors who oversee the newsroom 24 hours a day. The head and the head's team decide the order of articles on the

Table 1. Layout of Homepage with Control and Treatment

Control	Treatment
Human Editor	Algorithm
Slot 1: article 1	Slot 1: article 1
Slot 2: article 2	Slot 2: article 2
Slot 3: article 3	Slot 3: article 3
Slot 4: article 4	Slot 4: article 6
Slot 5: article 5	Slot 5: article 4
Slot 6: article 6	Slot 6: article 5
Slot 7: article 7	Slot 7: article 7

Notes. The table shows how the layout of the homepage of the website changes in the treatment with algorithmic recommendations relative to control with the human editor curating. In the example shown here, the algorithm moves the article on slot 6 upwards.

homepage. Outside of regular working hours, the deputy head takes the editorial decisions.

4.2. Field Experiment

Our field experiment took place when it was rare for major legacy news outlets to experiment with algorithmic curation. The *New York Times*, for instance, was among the first major outlets to experiment with personalized newsfeeds in June 2019. The experiment was run from December 2017 to May 2018 and included all visitors across desktop and mobile devices.

In the control condition, the human editor selects all articles on the homepage without any personalization for any user. In the treatment group, for each user, the algorithm personalizes the recommendation on slot 4, whereas the rest of the homepage remains the same. Table 1 provides a simple example where the recommendation algorithm moves up an article from slot 6 to slot 4. In our field experiment, the median recommendation moved up an article by 10 slots.

The recommendation system is based on a model put forward by Google engineers for Google News (Liu et al. 2010; see Online Appendix for technical details). The goal is to show users an article in their most preferred category to increase user clicks. Throughout this paper, we use clicks and engagement interchangeably. The algorithm's objective function is to get user i to click on the treatment slot at any time t . We identify users based on a unique cookie ID. The system predicts

individual click-through rates by category and then selects an article in the category with the highest predicted clickthrough rate from the pool of about 80 articles that the human editor has selected to appear on the homepage at any given moment. The median number of recommended article categories over an entire day is 255, creating sufficient opportunity for the algorithm to identify user preferences. The algorithm updates prediction scores daily, accounting for the available click history till then. If the system does not have enough data on a user, it assigns a recommendation based on current news trends derived from aggregate data. About 20% of the overall sample (about 56% of users) accounts for only one visit in the observed period. With more user engagement, recommendations are increasingly using personal data.

The randomization is at the user-session level, where a new session commences when the user is either inactive for 30 minutes or reloads the homepage. This level of randomization provides the statistical power to exploit variation within users. We can add user fixed effects to control for time-invariant unobserved heterogeneity (e.g., preferences) and isolate the effect of heterogeneity in the volume of personal data. Overall, the experiment involves a subtle treatment. This simplicity allows us to analyze reader behavior in a precise yet rich setting without disrupting news consumption on the site more generally. Slot 4 gets about 3% of the total clicks on the website (Table 2). Still, given the large overall traffic on the site and the fact that the experiment ran for multiple months, we have enough power to identify the heterogeneity in the relative performance of the human and the algorithms concerning the volume of personal data and variation in user preferences because of temporal changes in the types of news stories available.

To check the validity of our randomization procedure, we analyze the average assignment of individuals into treatment and control groups through the experiment, based on their pretreatment characteristics. We test the equality of means based on the percentage of days active before the experiment, the total number of clicks, clicks per day, clicks during work hours, and the geography of users across treatment

Table 2. Summary Statistics

	Mean	Standard deviation	Min	Max
Clicks on slot 4	0.0277	0.185	0	110
Total clicks	0.754	1.349	0	2809
Prior visits	2.805	2.178	0	11.72
Clicks on slot 4 (<i>New Year Bug</i>)	0.0242	0.165	0	54
High variance days	0.186	0.0389	0	1
Demand variance	3.172	0.950	1.304	6.150

and control conditions. The sample is well balanced across all observables (see Online Appendix Table A.2), indicating that the randomization has worked as intended.

During the experimental period, a coding bug impaired the ability of the algorithm to access daily updates of personal data. The engineers mistakenly hard coded the year to 2017 in the routine that pulls historical personal data. No personal data were available since the project began at the end of 2017. The bug remained unnoticed for the first six days of January 2018. During that period, the algorithm did not have access to fresh personal data and made recommendations in the first week of January 2018 using personal data from December 2017. The algorithm used increasingly outdated information as time passed during the bug period. If the flow of data is important, not using updated personal data should render recommendations increasingly less valuable to users. A user who visited the website on the 1st of January would see more “relevant” content than if they visited on the 6th of January. The mean number of clicks on slot 4 in a user session during the bug period is 0.024 (Table 2). Because the control group continued to see the human-curated version of the homepage, the coding bug isolates the effect of less recent personal data from other confounding factors, such as the news cycle in the first week of January 2018. As we will show, the relative performance of the recommendation algorithm is different in the week of the coding bug compared with all other weeks. Hence, we exclude this week from our primary estimation sample and use corresponding data only when we use the coding bug variation.

4.3. Econometric Model

Our baseline specification compares clicks between treatment and control:

Clicks_{ist} = α + δTreat_{ist} + γ_t + μ_i + ε_{ist}. (6)

The unit of observation is user *i* in session *s*. We define a session to include all clicks a user makes after arriving at the homepage until there is inactivity for 30 minutes or the user navigates again to the homepage. The dependent variable is the sum of clicks that originate from the treatment slot (*slot* = 4).¹ If the algorithm performs better than the human editor, we should expect the estimate of the treatment effect, δ, to be positive and statistically significant. Our preferred specification includes day fixed effects, γ_{*t*}, to control for events affecting all users, potentially through the news cycle and user-level fixed effects (μ_{*i*}). We cluster standard errors at the user level to account for the serial correlation of preferences. We extend this model by estimating heterogeneous treatment effects guided by our theoretical model, particularly concerning the volume and flow of personal data and variation in content availability.

5. Results

We first discuss our results of the relative performance of algorithmic recommendations and human curation and then turn to counterfactual computations of the optimal combination of both and discuss revenue implications.

5.1. Relative Performance of Algorithmic Recommendations

5.1.1. Effects of User Engagement. In column 1 of Table 3, we find that the treatment effect is positive on average. The magnitude of the average effect is 2.5% (0.0007 is the treatment effect, whereas the mean number of clicks on slot 4 is 0.0277; see Table 2). Whereas the average effect implies that the algorithm outperforms the human editor, it masks a significant amount of heterogeneity, which is crucial from a managerial standpoint. To better understand the heterogeneity in treatment effects, we analyze how the treatment effect varies with the amount of prior user *Engagement*. This is a direct test of Hypothesis 1, which states that the algorithm should outperform the human if there are sufficient individual-level personal data. We measure the amount of personal data available to the algorithm using the number of prior visits by an individual. With each visit (and subsequent clicks), a reader reveals preference by engaging with different types of content. In column 2, we interact the treatment indicator with the number of past visits (volume) and the natural logarithm of user *i*’s visits to the website from December 2017 up to the focal session. This specification finds a negative coefficient of *Treatment*. However, the positive coefficient of *Treatment* × *Engagement* shows that clicks to articles on the treatment slot increase with the information available to the algorithm. The average treatment effect for users without a prior visit history is negative (28% decline relative to the mean), which, in

Table 3. Performance of Algorithmic Recommendations Relative to Human Curation

	Clicks on slot 4		
	(1)	(2)	(3)
<i>Treatment</i>	0.0007*** (0.00004)	−0.0081*** (0.00008)	−0.0125*** (0.00014)
<i>Treatment</i> × <i>Engagement</i>		0.0030*** (0.00003)	0.0084*** (0.00013)
<i>Treatment</i> × <i>Engagement</i> ²			−0.0009*** (0.00002)
Day FE	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes
Observations	137,689,847	137,689,847	137,689,847

Notes. The dependent variable is the number of clicks on slot 4. The unit of observation is user-session. *Engagement* is the natural logarithm of user *i*’s prior visits to the homepage at time *t*. Standard errors clustered at the user level in parentheses. FE, fixed effects.

****p* < 0.001.

line with Hypothesis 1, indicates that the human editor can provide better recommendations relative to an algorithm that only has access to limited personal data. Table A.3 in the Online Appendix shows that at times when it is more likely that there is a less experienced editor in charge (i.e., night shifts and week-ends), the algorithm does better. Using a natural experiment of a change in lead editors, we show that the algorithm performs better after a younger, perhaps less experienced editor takes charge. For this analysis, we focus on the subsample where little personal data are available to the algorithm. Overall, these results validate the model predictions about the role of personal data and human expertise in such online settings. These heterogeneity results have clear managerial implications as well. When the algorithm's prior is not good enough, often because of a cold-start problem, an (experienced) human expert can curate better content. In our setting, we show that the human can outperform the algorithm in some cases even after we limit the severity of the cold-start problem using standard techniques developed in the computer science literature.

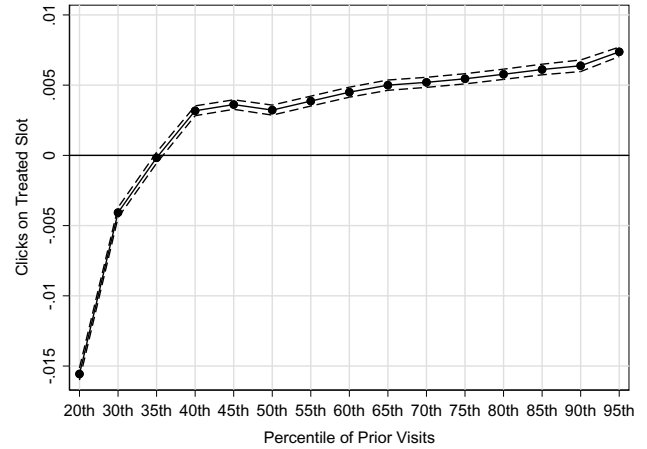
Next, as a test of Hypothesis 2, we look at a quadratic functional form for the prior volume of visits in column 3 of Table 3. We find that the quadratic term is negative, suggesting that additional information helps the algorithm but at a diminishing rate. We also look at finer bins based on the distribution of prior visits. We extend the model specification as follows:

$$\begin{aligned} Clicks_{is}^k = & \delta_1 Treat_{is} + \sum_q \delta_q (Treat_{is} \times PriorVisits_{qis}) \\ & + \lambda_q PriorVisits_{qis} + \gamma_\tau + \mu_i + \varepsilon_{is}, \end{aligned} \quad (7)$$

where $PriorVisits_{qis}$ indicates whether user i in session s is in the q th percentile of $PriorVisits$ of all users across all sessions.

In Figure 2, we plot $\hat{\delta}_1 + \hat{\delta}_q PriorVisits$ for 15 percentile bins of $PriorVisits$. The human editor outperforms the recommendation system when the algorithm has limited access to personal data. Our results show that the human editor has a comparative advantage up to the 35th percentile of personal data, which corresponds to when the algorithm has information from up to three visits per user. Around the threshold of the 40th percentile (five visits), the gap between algorithmic performance and human curation becomes positive and economically significant. With more available personal data, the effect of the algorithmic recommendations continues to increase, but at a decreasing rate. Beyond the level of the 65th percentile (50 visits), relative click-through rates stay at similar levels of economic significance. The effect size of algorithmic recommendation relative to human curation is about 18.5% for an individual with 50 prior visits.²

Figure 2. Heterogeneity in Relative Performance Because of (Prior) Engagement Volume



Notes. The figure plots the coefficients of the treatment effects associated with algorithmic recommendations relative to the human editor along with 99% confidence intervals based on the different bins specified in Equation (7). On the vertical axis, we have the magnitude of the treatment effects of algorithmic recommendations relative to the human. We have the percentiles of the number of visits of an individual user on the horizontal axis. The dependent variable is the number of clicks on slot 4. The unit of observation is user session.

Overall, in line with Hypothesis 2, we find that the relative efficacy of the algorithm improves with additional personal data but at a diminishing rate. The managerial implication is that leveraging the digital footprints of readers, especially at the early stages and despite the diminishing returns, can increase the probability of engagement. In our setting, about 56% of the users visit the website only once, so accumulating data on these individuals has a high payoff.

5.1.2. Stock vs. Flow of Personal Data. We have been able to compare the same user's behavior in the treatment group to their behavior in the control group, holding unobserved time-invariant heterogeneity fixed and controlling for the average effect of the amount of personal data available to the algorithm. However, as user engagement increases, the algorithm also gains information about user preferences. Each visit to a news website creates a *flow of data*, which, over time, accumulates to a *stock of data* on click behavior.

We can disentangle the value of the stock versus the flow of data for the algorithm using quasiexperimental variation from the coding bug described. In principle, the updating frequency of the algorithm is the same for all users. During the first week of January 2018, the engineering team missed the algorithm's failed attempts to pull individual-level data that weren't stored in the system. Hence, although users continued to visit the website, creating data on their preferences, the algorithm did not use this information. We can compare users with

the same level of engagement but different levels of available personal data. This natural experiment allows a direct test of Hypothesis 3. We can investigate how personal data flow affects algorithmic recommendations’ relative performance using this variation and test whether there is substitutability between the stock and flow of personal data.

In our analysis in columns 1 and 2 of Table 4, we focus on a sample that covers December 2017 and January 2018. We define the indicator *New Year Bug* equal to one on January 1–6, 2018, when the bug in the code was in effect or zero otherwise. We focus our attention on the interaction term of whether a user was in the treatment group and whether we observe the user on the website during the *New Year Bug* period. Column 1 shows that clicks on slot 4 are significantly lower. The total effect even turns negative and corresponds to a reduction of 9.5%. In column 2, we introduce a linear trend capturing each successive day that went by with the bug as part of the code pulling personal data. We find that the interaction term between the treatment and the linear trend variable is also significantly negative, implying that users in the treatment group click less with each additional day without updated data. Finally, we assess whether the stock of existing data can mitigate this reduced precision in personalization. We created a variable that captures the number of visits an individual had before January 1, 2018, as the stock and analyzed its interaction with the treatment, focusing only on the six days in which the bug was in effect. The result in column 3 of Table 4 shows that the interaction coefficient is positive. This positive coefficient suggests, in line with Hypothesis 3, the stock of data can

help predict user preferences and mitigate the need for continuous updating.

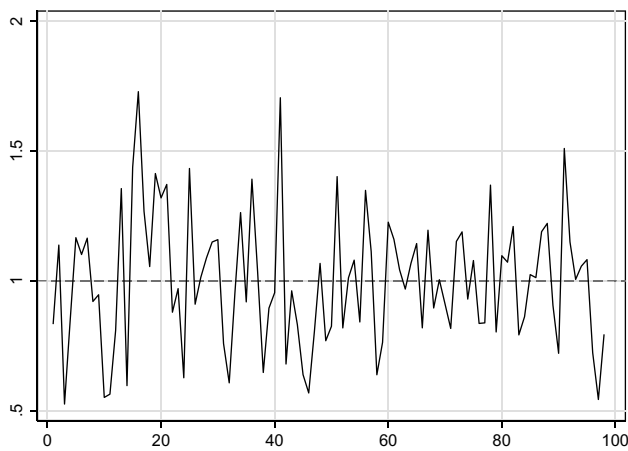
5.1.3. Effects of Variation in Content on Relative Efficacy of Automated Recommendations. We now explore whether the relative performance of automated recommendations depends on the variation in content, which impacts reader interest in different types of content on a day-to-day basis. The intuition is that as the context or external events change, they can affect reading preferences, albeit temporarily. In such scenarios where ground truth changes, it is possible that the human might be able to apply broad knowledge better than the algorithm. We test this intuition behind Hypothesis 4 using two measures of demand variance. We aggregate total daily clicks in the 10 most popular categories for the entire observed period and calculate the standard deviation of total clicks across all days for each category. We then compute the ratio of daily total clicks of a category and the category’s standard deviation, which gives us a measure of the demand variation of each category on a given day. Based on this, we create two variables that characterize how attention grabbing the news stories in a given category of a given day are. First, we define *HighVarianceDays* as days where at least one category’s demand variation measure is larger than the 95th percentile. In our sample, 20% of the days are *HighVarianceDays*. Second, we define the continuous variable *DemandVariance* to be equal to the maximum demand variance of any of the categories. The average day has a maximum demand variance of 3.10, but there is substantial variation (standard deviation of 1.16). The fifth percentile is 1.74, whereas the 95th percentile is 5.06. Figure 3 illustrates the variation in

Table 4. Stock vs. Flow of Personal Data: Impact on Automated Recommendations Because of Coding Bug

	Clicks on slot 4		
	(1) Dec–Jan	(2) Dec–Jan	(3) Bug period
<i>Treatment</i>	0.0020*** (0.00006)	0.0020*** (0.00006)	−0.0007 (0.00057)
<i>Treatment</i> × <i>New Year Bug</i>	−0.0046*** (0.00014)		
<i>Treatment</i> × <i>New Year Bug Day Trend</i>		−0.0014*** (0.00003)	
<i>Treatment</i> × <i>Visits Prior to New Year Bug</i>			0.0006*** (0.00015)
Day FE	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes
Observations	71,901,149	71,901,149	5,907,837

Notes. The dependent variable is the number of clicks on slot 4. The unit of observation is user session. *Treatment* indicates whether user *i* is in the treatment group in session *s*. *New Year Bug* indicates the first week of January 2018. *New Year Bug Day Trend* is a linear trend for the first week of January 2018. *Visits Prior to New Year Bug* is the natural logarithm of user *i*’s prior visits to the homepage before the first week of January 2018. Standard errors are clustered at the user level in parentheses. FE, fixed effects.

****p* < 0.001.

Figure 3. Empirical Demand Variance

Notes. The figure plots *demand variation* for each day in our sample relative to the monthly mean. For example, a value of 1.5 means that there was 50% more demand variation on this day than on average in the corresponding month. The horizontal axis depicts the days in our sample.

demand variance we exploit. We plot *DemandVariance* for each day in our sample relative to the monthly mean. Throughout our observation period, we observe days with substantially less demand variance than average and days with substantially more demand variance than average. In columns 1 and 2 of Table 5, we show that the treatment effect is significantly smaller on days where total clicks are more volatile. The estimates suggest that the treatment effect is about two times smaller on days with particularly high demand variance. Similarly, a one-standard-deviation increase in *DemandVariance* leads to a 27% lower treatment effect. These results imply that reverting to human expertise

Table 5. Results: Content Variation

	Clicks on slot 4	
	(1)	(2)
<i>Treatment</i>	0.0011*** (0.00004)	0.0052*** (0.00012)
<i>Treatment</i> × <i>HighVarianceDays</i>	−0.0021*** (0.00009)	
<i>Treatment</i> × <i>DemandVariance</i>		−0.0014*** (0.00004)
Day FE	Yes	Yes
Individual FE	Yes	Yes
Observations	137,689,847	137,689,847

Notes. The dependent variable is the number of clicks on slot 4. The unit of observation is user session. *Treatment* indicates whether user i is in the treatment group in session s . *HighVarianceDays* indicates days where at least one category's demand variation measure is larger than the 95th percentile. *DemandVariance* is the largest value of any category in mean clicks per category and day as a proportion of the standard deviation of clicks per category across the entire sample. Standard errors clustered at the user level in parentheses. FE, fixed effects.

*** $p < 0.001$.

can lead to better performance when there is variation in preferences over time. In line with Hypothesis 4, managers need to account for the limitations of algorithms trained within narrow contexts.

5.2. Optimal Combination of Human Curation and Algorithmic Recommendation

Given the results we have discussed, a promising strategy could be to use a weighted combination of human curation and algorithmic recommendation. Our results suggest that the editor can better judge the inherent relevance of news articles of a particular day than the recommendation algorithm in the absence of sufficient personal data. Hence, whatever information an editor uses for this judgment is not a direct input into the algorithm deployed within our field experiment. Again, in the absence of personal data, the algorithm uses aggregate data on news demand to solve its cold-start problem.

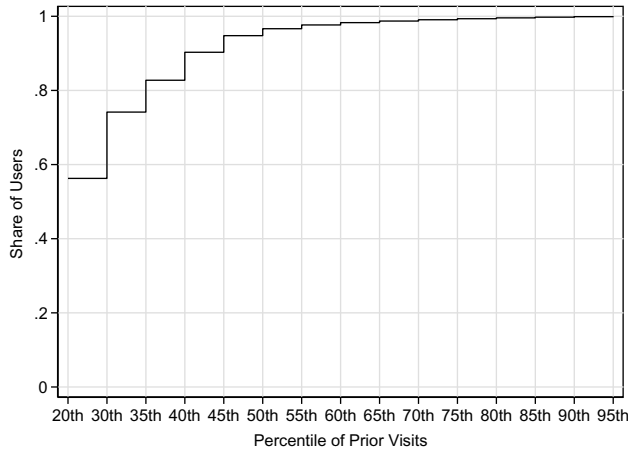
We take this as a motivation to use our estimates to calibrate a set of alternative scenarios. We calculate by how much total clicks would change if NEWS implemented algorithmic recommendation throughout and by how much total clicks would change if they implemented the optimal combination of algorithmic recommendation and human curation. We then engage in a (partial-equilibrium) back-of-the-envelope calculation of the revenue implications of either policy, taking implementation costs (that is, labor costs) into consideration.

5.2.1. Alternative (Counterfactual) Scenarios. Translating insights from the user-session level to the firm level demands some caution. As suggested in Hypothesis 5, with sufficient heterogeneity among users, what we see at the disaggregated level might be different at the aggregate level (fallacy of composition).

User engagement typically follows a skewed distribution (Simonov and Rao 2022), and the same is true in our case. Figure 4 shows that relatively few users account for high volumes of engagement. Only about 10% of users show engagement larger than the 50th percentile. Relating the distribution of users and engagement volume to a break-even point at the 35th percentile of user data (see Figure 2), automated recommendation has a lower performance relative to human curation for about 56.5% of users.

This skewed user engagement distribution highlights important heterogeneity that could influence technology adoption decisions. We run a counterfactual computation using the observed data from the control group and our treatment effect estimates to quantify the benefits of an optimal combination of human curation and algorithmic recommendation. We analyze three potential strategies news outlets could employ while serving their users. In particular, we analyze whether NEWS should go back to only human curation, roll out the algorithmic recommendations to all users and entirely forgo human curation, or use a

Figure 4. Cumulative Distribution of Users by Number of (Prior) Visits



Notes. The figure plots the cumulative distribution of users according to their visits (total engagement). We suppress the values before the 20th percentile on the horizontal axis because of limited variation in a user's prior visits. The median number of visits is 16.

combination of human curation and algorithmic recommendation. Relatedly, we can ask whether—*ceteris paribus*—it would be better to implement a decision rule on when to use human curation and when to use algorithmic recommendation based on the volume of engagement. In particular, we calculate the average monthly total clicks for different scenarios:

$$\widehat{TotalClicks}_H^k = \frac{1}{M} \sum \frac{1}{\theta} \sum_q \overline{Clicks}_{qm}^{ck} N_{qm}^c \quad (8)$$

$$\widehat{TotalClicks}_A^k = \frac{1}{M} \sum \frac{1}{\theta} \sum_q (\overline{Clicks}_{qm}^{ck} + \hat{\delta}_q) N_{qm}^c \quad (9)$$

$$\begin{aligned} \widehat{TotalClicks}_{HA}^k &= \frac{1}{M} \sum \begin{cases} \frac{1}{\theta} \sum_q (\overline{Clicks}_{qm}^{ck} + \hat{\delta}_q) N_{qm}^c, & \text{if } \hat{\delta}_q > 0 \\ \frac{1}{\theta} \sum_q \overline{Clicks}_{qm}^{ck} N_{qm}^c, & \text{if } \hat{\delta}_q \leq 0 \end{cases} \quad (10) \end{aligned}$$

where M is the total number of months, θ is the share of users in the control group, and $\overline{Clicks}_{qm}^{ck}$ is the average number of clicks on slot k of control group users in percentile q in terms of exposure to content volume or prior visits to the homepage in month m . The monthly number of users in percentile q in the control group is denoted by N_{qm}^c . The $\hat{\delta}_q$ is based on the estimates from Equation 7 plotted in Figure 2. By design, we can estimate total clicks only on the treatment slot. However, we can also extrapolate to total clicks on any slot of the entire homepage under some simplifying assumptions. The main assumption we make is that, at scale, the treatment effect is constant across slots. A limitation of the analysis is that we abstract away from the possibility that clicks on one article can cannibalize clicks on other articles.

Table 6 shows that using only algorithmic recommendation (when holding engagement volume at the average) would lead to a decrease of 1% in clicks on the treatment slot. Moreover, under the assumption of constant treatment effects, this would add up to about 79,800 fewer total clicks when applied to all slots on the homepage. However, we can expect an increase in total clicks of 13% in a scenario of human curation when engagement volume is below the threshold of the 35th percentile (see Figure 2) and algorithmic recommendation when it is above that threshold. When we simulate total clicks on the entire homepage, we get an estimate of an additional 4.5 million clicks.

In sum, we can conclude that a hybrid model with both algorithmic recommendations and human curation strongly dominates implementing only algorithmic recommendations or using only human curation. These results support the notion of Hypothesis 5. They are in line with a study of algorithmic recommendations in the Norwegian news industry, which concludes that many outlets have found it undesirable to fully automate recommendations and look for approaches that combine automatic recommendations with editorial choices (Gulla et al. 2021).

5.2.2. Monetary Implications. Under a few simplifying assumptions, we can use the simulation results to

Table 6. Alternative Recommendation Technology Scenarios

	Algorithm only (1)	Human and Algorithm (2)
Relative to human only	−1.0% (−2.3%, 0.4%)	13.1% (12.2%, 14.0%)
Additional clicks (slot = 4)	−11,685	169,886
Additional clicks (entire homepage)	−79,846	4,562,972

Notes. The *relative* effect is the monthly average total clicks in either the algorithm-only (Equation 9) or human-and-algorithm scenario (Equation 10) in relation to total clicks in the human-only scenario (Equation 8). Values in brackets indicate 99% confidence bands based on the standard errors of the treatment effect estimate.

speculate how total clicks would convert into monetary units and relate that to the implementation cost of the hybrid recommendation system.

We start with the assumption that clicks are proportional to advertising revenues. In scenario 1 (implementation on one slot), we expect a hybrid recommendation system to lead to an additional 170,000 clicks. When implemented on the entire homepage (scenario 2), we expect 4.5 million additional clicks (Table 6). According to industry reports, the average click-through rate on display is about 0.35%, which implies total additional monthly clicks on a particular ad would be 595 and 15,750. With two prominent ads on each page, the total clicks on ads would be about 1,200 and 31,500, respectively.³ The average cost per click is about \$0.58, which implies that the total additional monthly revenues accruing to the news outlet are \$500 and \$18,270, respectively.

In the short run, we can assume that the labor costs of human editors are sunk; hence, we only need to consider the labor costs of setting up an automated recommendation system. The average monthly salary of a data scientist in Germany, along with benefits, comes to a total of approximately \$7,200.⁴ Hence, for the hybrid recommendation system to break even, it has to scale up to the entire homepage, not only on the one treatment slot we used in our field experiment.

Assuming that the hybrid system is implemented for the entire homepage, a news outlet would need a total traffic of about 47.5 million clicks to compensate for one month of implementation costs of one full-time data scientist working on the project.⁵ The assumption till now is that it takes one month's work from a data scientist to set up the recommender system with zero ongoing variable costs. This is in line with the fundamental premise of automation, where the fixed cost is high and the variable costs are low. We can relax this assumption and analyze different scenarios. For example, if the implementation takes six months, the outlet would need 285 million clicks to break even, assuming no additional variable costs going forward.

This exercise aims to demonstrate how to use these estimates to evaluate alternative scenarios for news outlets with different audience sizes (local versus national), the number of ads on a page, and algorithmic performance. Of course, a caveat is that these estimates are partial equilibrium because such a change could also free up the curating human editors to carry out other tasks.

6. Discussion

6.1. Internal Validity and Robustness

We carry out several checks to ensure the robustness of our results. Table A.4 in the Online Appendix shows that our baseline results hold with alternative econometric models (log clicks, linear probability model (LPM), Poisson, and negative binomial). In Table A.5 in

the Online Appendix, we show that our results are robust to alternative fixed effects (user-week, user-day, user-hour, user-hour-of-the-day), suggesting that there are no differences in reading behavior because of the news cycle or related to different preferences or strategies of human editors. Next, in columns 1–3 of Table A.6 of the Online Appendix, we show that the level of randomization (session level) is not the driver of our results because restricting our sample to simply the first session of the hour (column 1), the first session of the morning (column 2), and the first session of the afternoon (column 3) leaves the treatment effect qualitatively unchanged.

A possible improvement of the recommendation algorithm could be to take signals of the importance of individual articles into consideration, based on the ranking of an article that the human editor has decided. We look at snapshots of data of the homepage layout at five-minute intervals for several days during the experimental period. A caveat is that the analysis will be correlational because we can only exploit the variation within the treatment group over time. Here, a human editor choosing a higher numerical rank implies lower importance. For instance, an article with a rank of 24 is considered of lower importance, as determined by the expert, because it will be lower down the page, relative to an article with a rank of 14. We show that results in columns 4 and 5 of Table A.6 in the Online Appendix are in line with intuition. In column 4, we can see that, on average, an article that has a higher numerical rank in the control condition (a higher number) will lead to fewer clicks when moved to slot 4 in the treatment condition. The result in column 4 is in line with the editor placing an article lower down the page because it is less relevant for an average reader. The interaction of personal data with rank leads to a higher number of clicks, implying that as the algorithm learns more about the individual, it can match preferences better, leading to higher engagement. In column 5, we restrict the articles to those that have a numerical rank of 14 or higher (at least 10 ranks higher than the treated slot) to ensure that the “pulling up” of articles will be meaningful, in that the readers might not have seen the article otherwise. The results are qualitatively similar to those in column 4, implying that an article that has a lower probability of being clicked on when it was lower down the page (higher rank). Additionally, the article is more likely to be clicked when ranked lower by the algorithm. Combined with a better fit (because of more prior data) leads to more clicks on the treated slot.

6.2. External Validity and Implications

Our study has implications for a variety of different contexts that contemplate the adoption of algorithmic recommendation technologies.

First, our result in the case of limited volume of user-level engagement suggests that human expertise is best suited for relatively complex situations with limited information for the algorithm. Indeed, in a different setting (finance), human traders outperform algorithmic traders in less routine trades (because of limited information), implying that the optimal strategy is a combination of the two (Brogaard et al. 2021). More generally, the potential of the human-algorithm combination is consistent with the broad idea that humans can have an important role when subjective decisions are taken in a predictive context (Agrawal et al. 2018). Human expert decisions based on accumulated knowledge (captured by the human's prior μ in the model) are not simple to replicate algorithmically (Wang et al. 2017). The managerial implication of our results is that human intervention should occur when the algorithm's prior is not good enough, which is one form of a cold-start problem. The computer science literature has focused on solving this problem (Gope and Jain 2017, Sethi and Mehrotra 2021) in different contexts. A natural alternative could be reverting to the status quo of human expertise even after mitigating the severity of the cold-start problem by providing aggregate data to the algorithm as recommended by the literature. Moreover, our results show that the efficacy of a human expert does vary with experience and should be taken into account in other contexts as well. In the setting of online chess, humans with greater expertise do perform at higher levels (when measured against the algorithm) (Miric et al. 2020)—so do salespeople in sales force management (Hu et al. 2022). Indeed, this could explain why platforms such as HBO Max and Netflix are bringing back human curation (Alexander 2019) to mimic the strategy of companies such as Spotify that have been using human expertise along with algorithmic recommendations (Aguilar et al. 2021).

Next, our result on variation in preferences suggests that human expertise can be more flexible in adapting to external environmental changes in content availability. Indeed, this pain point for algorithms carries over into recommendations in video games (Wang et al. 2022) in flash crashes in the stock market (Jurich et al. 2020). The performance of ML models when there is significant variation in user preferences at a point in time or over time has been a cause for concern in terms of their practical applications, known as time drift or concept drift in the literature (Rafieian and Yoganarasimhan 2022). Variation in preferences has significant practical implications for companies because external events impact users daily and affect the demand for products. Relatedly, as another example, the inability of the algorithm to account for variability in external conditions in the real estate market is said to have played a significant role in the failure of Zillow's iBuyer

in 2021 (Stokel-Walker 2021). The limits of the viability of algorithmic predictions were reached when the underlying model did not account for the shortage of contractors, inflation expectations, and changing housing demand. Our study demonstrates that human intervention using broad knowledge of the context should override the status quo or business-as-usual algorithmic recommendations (i.e., the combination of human and algorithm we propose). Our theoretical and empirical results show that algorithms trained on data within a specific context are not flexible enough to deal with constant changes in the ground truth.

Finally, the stock versus flow result captures the dynamics in which the algorithm, because of a technical glitch, ends up in a unique scenario of making a bad recommendation. Whereas a firm would never willfully experiment with a bad recommendation, such system glitches are common (Heath 2022, Mehta and Singh 2022) and are analyzed in academic contexts through online platforms such as eBay (Hui and Liu 2022) and omnichannel retailers (Narang et al. 2020). This suggests that human recommendations could substitute for bad automated recommendations, similar to a setting where human expertise is used to flag harmful content to ensure that the algorithms do not amplify bad recommendations (Scott and Kayali 2020). These examples and our (theoretical and empirical) results demonstrate the trade-offs of using an automated model trained on detailed data within narrow contexts. System glitches can hinder performance when the algorithm cannot access a continuous data feed. Human experts use their broad-based experience to curate content that can be a good alternative to automation in such settings.

7. Conclusion

We run a field experiment to quantify the scenarios under which a company should adopt an algorithmic recommendation technology relative to human curation in the context of online news. We propose a theoretical framework where human experts use their experience to curate articles but cannot personalize at scale and contrast it to scalable recommendations by an algorithm trained on detailed but narrow personal data.

Our empirical analysis demonstrates that automated recommendation can outperform human curation on average. However, this result depends on the experience of human editors, the amount of personal data available to the algorithm, and variation in the external environment causing variation in demand.

Our results are not without limitations. Our field experiment is narrow because it only tests how one algorithm performs relative to (a particular set of) human editors. On the flip side, the strength of the analysis is that we can fix the algorithm and analyze variation in

data inputs. We still believe that our theoretical and empirical results are useful to analyze similar contexts, as discussed, because an extensive literature review of the related technical literature demonstrates that the (offline) performance of various algorithms remains very similar (Wu et al. 2022). Indeed, the performance of certain (older) algorithms persists over time relative to newer, more computationally intensive innovations.

The context we analyze theoretically and empirically comprises an ad-funded business model. Indeed, the objective function of our field partner is to simply maximize clicks. Our results may apply only to ad-supported news media. Outlets with a subscription model would maximize the number of (recurring) visitors, whereas an ad-funded one can focus purely on the overall number of clicks.

Personalized recommendation is a common form of ML on online platforms (Lee and Hosanagar 2019, Schaefer and Sapi 2020). We have also discussed other examples in finance and real estate and the role of human expertise in the presence of algorithmic recommendations. Our results add to an interdisciplinary literature that provides consistent evidence for complementarity between humans and algorithms. More research on these issues is imperative to generate regularities across a larger number of contexts.

Acknowledgments

The authors thank Tobias Gesche, Miguel Godinho de Matos, Tobias Kretschmer, Anja Lambrecht, Dokyun Lee, Klaus Miller, Lorien Sabatino, Tianshu Sun, Huseyin Tanriverdi, Catherine Tucker, Tommaso Valletti, Dogukan Yilmaz, as well as conference and seminar participants at KOF-ETH Zurich, MIT CODE, Media Economics Workshop (Lecce), MIT IDE Annual Conference, Marketing Science (Rome), SCECR (Hong Kong), Munich Beliefs about Society and Politics Workshop, Workshop on the Economics of Data and AI at the European Commission, USC Marshall, Santa Clara University, University of Houston, CLE-ETH Zurich, HEC Lausanne, KU Leuven, University of Bologna, ZEW Mannheim, CIST, Paris Seminar on the Economics of Digitization, Zurich Media and Digitization Workshop, NBER IT and Digitization Spring Meeting, NBER IO Spring Meeting, Digital Platforms Seminar hosted by Toulouse School of Economics, Harvard Business School Digital Initiative Seminar, UT Austin McCombs School of Business, and CESifo Economics of Digitization Conference for helpful suggestions.

Author order of the first two authors, who contributed equally, is certified random using the AEA author randomization tool and publicly archived under as_5-qIu16iQ.

The study is registered at the AEA RCT Registry under AEARCTR-0004264.

Endnotes

¹ As a robustness check, we also analyze our baseline results using the logarithm of clicks, the probability of any click as well as other nonlinear models in Table A.4 in the Online Appendix.

² From an identification perspective, we note that prior treatment status does not predict revisits to the website. This is because we have a subtle treatment involving manipulating slot 4 which accounts for less than 3% of the engagement. In robustness checks, we demonstrate that the treatment effect is consistent across different specifications and use of fixed effects. See Section 6.1 for more details.

³ See <https://blog.hubspot.com/agency/google-adwords-benchmark-data> for an overview of the industry numbers and the specific values we use in these computations. Note that aggregate data on click-through rates include about a third of users with ad blockers installed.

⁴ See https://de.glassdoor.ch/GehC3A4lter/germany-data-scientist-gehalt-SRCH_IL.0,7_IN96_KO8,22.htm?countryRedirect=true and https://www.destatis.de/EN/Themes/Labour/Labour-Costs-Non-Wage-Costs/_node.html for details on salary estimates and non-wage benefits across industries and jobs.

⁵ These traffic numbers match the number of visits to Germany's top 25 news outlets in a month. See <https://www.ivw.de/englische-version> for details on these numbers. For comparison, monthly clicks to the *Wall Street Journal*, *Los Angeles Times*, and *Boston Globe* are around 120 million, 63 million, and 30 million, respectively.

References

- Agrawal A, Gans J, Goldfarb A (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence* (Harvard Business Press, Cambridge, MA).
- Aguiar L, Waldfogel J (2021) Platforms, power, and promotion: Evidence from Spotify playlists. *J. Indust. Econom.* 69(3):653–691.
- Aguiar L, Waldfogel J, Waldfogel S (2021) Playlisting favorites: Measuring platform bias in the music industry. *Internat. J. Indust. Organ.* 78:102765.
- Alexander J (2019) HBO launches 'recommended by humans' tool to help you escape algorithm nightmares. *Verge* (August 6), <https://www.theverge.com/2019/8/6/20757370/hbo-recommendation-algorithm-netflix-streaming-euphoria-game-of-thrones-succession>.
- Amatriain X, Lathia N, Pujol JM, Kwak H, Oliver N (2009) The wisdom of the few: A collaborative filtering approach based on expert opinions from the web. *Proc. 32nd Intern. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 532–539.
- Bodó B (2019) Selling news to audiences—a qualitative inquiry into the emerging logics of algorithmic news personalization in European quality news media. *Digital Journalism* 7(8):1054–1075.
- Brogaard J, Ringgenberg MC, Roesch D (2021) Does floor trading matter? *J. Finance* Forthcoming.
- Chiou L, Tucker C (2017) Search engines and data retention: Implications for privacy and antitrust. Working paper, National Bureau of Economic Research, Cambridge, MA.
- Choi S, Kim N, Kim J, Kang H (2022) How does artificial intelligence improve human decision-making? Evidence from the AI-powered Go program. USC Marshall School of Business Research Paper, USC Marshall School of Business, Los Angeles.
- Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management J.* 41:1381–1411.
- Cowgill B (2018) Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Working paper, Columbia University, New York.
- DeGroot MH (1970) *Optimal Statistical Decisions* (Wiley, Hoboken, NJ).
- Gentzkow M, Shapiro JM (2010) What drives media slant? Evidence from US daily newspapers. *Econometrica* 78(1):35–71.
- Goldfarb A, Tucker CE (2014) Conducting research with quasi-experiments: A guide for marketers. Rotman School of Management Working Paper No. 2420920, Rotman School of Management, Toronto.
- Gope J, Jain SK (2017) A survey on solving cold start problem in recommender systems. *2017 Intern. Conf. Comput. Comm. Automation (ICCCA)* (IEEE, Piscataway, NJ), 133–138.

- Gulla J, Svendsen R, Zhang L, Stenbom A, Frøland J (2021) Recommending news in traditional media companies. *AI Mag.* 42(3):55–69.
- Heath A (2022) A Facebook bug led to increased views of harmful content over six months. *Verge* (March 31), <https://www.theverge.com/2022/3/31/23004326/facebook-news-feed-downranking-integrity-bug>.
- Hu S, Zhang J, Zhu Y (2022) Zero to one: Sales prospecting with augmented recommendation. MIT Sloan Research Paper 6492-20. Preprint, submitted January 12, <http://dx.doi.org/10.2139/ssrn.4006841>.
- Hui X, Liu M (2022) Quality certificates alleviate consumer aversion to sponsored search advertising. Preprint, submitted July 7, <http://dx.doi.org/10.2139/ssrn.4155772>.
- Jannach D, Jugovac M (2019) Measuring the business value of recommender systems. *ACM Trans. Management Inform. Systems* 10(4):1–23.
- Jurich SN, Mishra AK, Parikh B (2020) Indecisive algos: Do limit order revisions increase market load? *J. Behav. Experiment. Finance* 28:100408.
- Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9):100804.
- Krishnan V, Narayanashetty PK, Nathan M, Davies RT, Konstan JA (2008) Who predicts better? Results from an online study comparing humans and an online recommender system. *Proc. 2008 ACM Conf. Recommender Systems* (ACM, New York), 211–218.
- Lee D, Hosanagar K (2019) How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment. *Inform. Systems Res.* 30(1):239–259.
- Liu J, Dolan P, Pedersen ER (2010) Personalized news recommendation based on click behavior. *Proc. 15th Internat. Conf. Intelligent User Interfaces* (ACM, New York), 31–40.
- McCombs ME, Shaw DL (1972) The agenda-setting function of mass media. *Public Opinion Quart.* 36(2):176–187.
- Mehta I, Singh M (2022) Facebook fixes strange bug that spammed everyone's feed with celebrity fan posts. *TechCrunch* (August 24), <https://techcrunch.com/2022/08/24/strange-facebook-bug-is-spamming-everyones-feed-with-celebrity-page-posts/#:~:text=Facebook%20fixes%20strange%20bug%20that%20spammed%20everyone's%20feed%20with%20celebrity%20fan%20posts,-Ivan%20Mehta%2C%20Manish&text=Update%20August%2024%2C%202022%2C%206,has%20now%20fixed%20the%20error.>
- Miric M, Lu J, Teodoridis F (2020) Decision-making skills in an AI world: Lessons from online chess. Preprint, submitted February 15, <http://dx.doi.org/10.2139/ssrn.3538840>.
- Narang U, Shankar V, Narayanan S (2020) How does mobile app failure affect purchases in online and offline channels? Working Paper No. 4053, Stanford Graduate School of Business, Stanford, CA.
- Petrova M, Sen A, Yildirim P (2021) Social media and political contributions: The impact of new technology on political competition. *Management Sci.* 67(5):2997–3021.
- Peukert C, Reimers I (2018) Digital disintermediation and efficiency in the market for ideas. CESifo Working Paper Series No. 6880, Center for Economic Studies, Munich, Germany.
- Rafieian O, Yoganarasimhan H (2022) AI and personalization. Preprint, submitted May 30, <http://dx.doi.org/10.2139/ssrn.4123356>.
- Schaefer M, Sapi G (2020) Learning from data and network effects: The example of internet search. DIW Berlin Discussion Paper 1894, DIW Berlin, Berlin.
- Scott M, Kayali L (2020) What happened when humans stopped managing social media content. *Politico* (October 21), <https://www.politico.eu/article/facebook-content-moderation-automation/#:~:text=Pro%20Free%20From-,What%20happened%20when%20humans%20stopped%20managing%20social%20media%20content,harmful%20content%20from%20their%20platforms.&text=This%20article%20is%20part%20of,report%2C%20The%20Essential%20Tech%20Worker.>
- Sen A, Yildirim P (2015) Clicks bias in editorial decisions: How does popularity shape online news coverage? Preprint, submitted June 16, <http://dx.doi.org/10.2139/ssrn.2619440>.
- Senecal S, Nantel J (2004) The influence of online product recommendations on consumers' online choices. *J. Retailing* 80(2):159–169.
- Sethi R, Mehrotra M (2021) Cold start in recommender systems—A survey from domain perspective. Hemanth J, Bestak R, Chen JIZ, eds. *Intelligent Data Communication Technologies and Internet of Things* (Springer, Singapore), 223–232.
- Simonov A, Rao J (2022) Demand for online news under government control: Evidence from Russia. *J. Political Econom.* 130(2):259–309.
- Sinha R, Swearingen K (2011) Comparing recommendations made by online systems and friends. Accessed January 19, 2023, <https://www.ercim.eu/publication/ws-proceedings/DelNoe02/RashmiSinha.pdf>.
- Stokel-Walker C (2021) Why Zillow couldn't make algorithmic house pricing work. *Wired* (November 11), <https://www.wired.com/story/zillow-ibuyer-real-estate/#:~:text=Piskorski%20acknowledges%20that%20inflation%20concerns,bounds%20of%20its%20algorithm's%20ability.>
- Sun T, Yuan Z, Li C, Zhang K, Xu J (2023) The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy. *Management Sci.* ePub ahead of print June 8, <https://doi.org/10.1287/mnsc.2023.4828>.
- Valavi E, Hestness J, Ardalani N, Iansiti M (2022) Time and the value of data. arXiv preprint arXiv:2203.09118.
- Wang X, Yu L, Ren K, Tao G, Zhang W, Yu Y, Wang J (2017) Dynamic attention deep model for article recommendation by learning human editors' demonstration. *Proc. 23rd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 2051–2059.
- Wang Z, Yin J, Feng Y, Liu Y (2022) Modeling behavioral dynamics in digital content consumption: An attention-based neural point process approach with applications in video games. *Marketing Sci.* Forthcoming.
- Wu C, Wu F, Huang Y, Xie X (2022) Personalized news recommendation: Methods and challenges. *ACM Trans. Inform. Systems* 41(1):1–50.
- Yeomans M, Shah A, Mullainathan S, Kleinberg J (2019) Making sense of recommendations. *J. Behav. Decis. Making.* 32(4):403–414.