The Editor and the Algorithm: Returns to Data and Externalities in Online News[*]

Jörg Claussen
LMU Munich and CBS[†]

Christian Peukert
HEC Lausanne[‡]

Ananya Sen
Carnegie Mellon University[§]

January 11, 2021

Abstract

We run a field experiment to quantify the economic returns to data and informational externalities associated with algorithmic recommendation relative to human curation in the context of online news. Our baseline results show that algorithmic recommendation can outperform human curation in terms of user engagement. We find significant heterogeneity with regards to the *type of data* (personal data is much more effective than aggregate data, but diminishing returns set in quickly), *user types* (users closer to average consumption patterns engage more), *local data network effects* (the size of the user base increases engagement only for users close to average consumption patterns), and *data freshness* (missing updating of personal data quickly deteriorates performance). We characterize circumstances where human curation leads to higher engagement than algorithmic recommendation to highlight complementarities between the human editor and the algorithm. Investigating informational externalities highlights that personalized recommendation reduces consumption diversity and that these are reinforced over time. Moreover, users associated with lower levels of digital literacy and more extreme political views engage more with algorithmic recommendations.

[*] The study is registered at the AEA RCT Registry under AEARCTR-0004264. We thank Tobias Gesche, Miguel Godinho de Matos, Anja Lambrecht, Dokyun Lee, Klaus Miller, Lorien Sabatino, Tianshu Sun, Tommaso Valletti, 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, ZEW Mannheim, CIST, Paris Seminar on the Economics of Digitization, Zurich Media and Digitization Workshop, NBER IT and Digitization Spring Meeting, Digital Platforms Seminar and CESifo Economics of Digitization Conference for helpful suggestions.

[†] LMU Munich, Munich School of Management (ISTO) and Copenhagen Business School, j.claussen@lmu.de.

[‡] HEC Lausanne, Department of Strategy, Globalization and Society, christian.peukert@unil.ch.

[§] Carnegie Mellon University, ananyase@andrew.cmu.edu.

1 Introduction

Digitization has disrupted and led to innovation in equal measure in a wide variety of industries over the past two decades (Goldfarb and Tucker, 2019; Waldfogel, 2017). Going forward, Artificial Intelligence (AI) and Machine Learning (ML) technologies are expected to play an increasingly important role in the digital economy. Data is a critical factor in these technologies, and therefore key to enabling valuable innovation leading to productivity gains. The news industry is a prime example of an industry that was disrupted due to digitization and has identified ML's potential to shore up firm revenues by predicting individual level customer preferences and establishing trends in large proprietary datasets. In particular, there is a debate within the news industry whether human editorial decisions should be aided or replaced by algorithmic recommendations which can personalize at scale (Nicas, 2018; Isaac, 2019). These editorial decisions are crucial for reader engagement and necessarily involve subjective judgments about the 'newsworthiness' of stories. The news industry itself is rooted in a particularly dynamic environment, characterized by a continuous flow of new information, both on the supply-side and the demand-side. In such a setting, the circumstances under which predictive algorithms can aid human workers who have extensive domain expertise are mostly unclear.

More generally, the question of whether algorithms can aid human experts – depending on the type and diversity of data that machines are learning from – has important managerial implications. It can also inform the policy debates on consumer privacy issues and the market power of digital platforms. In the context of online news, there are additional concerns related to the political economy effects of algorithmic news curation. In particular, algorithmic recommendations might lead to unintended consequences and a (socially) less desired outcome if news platforms do not account for informational externalities of their readers. This assumes greater significance if readers confine themselves into echo chambers with algorithms trained on prior individual-level data reinforcing this phenomenon (Gentzkow, 2018).

We carry out a large-scale field experiment with a major news outlet in Germany to provide empirical evidence on the interrelated issues of algorithmic recommendation versus human curation, economic returns to data, and consumption diversity. At any time, the homepage of the news outlet's website features N articles. In the experiment, every time a user visits the homepage, she is randomly assigned to a control or treatment condition. In the control condition, a human editor chooses the order of the articles on the homepage. This order is the same for every user at a given point in time. In the treatment condition, the order is customized using algorithmic recommendation. While the first three articles remain the same as in the control condition, the algorithm may replace the editor's choice of the fourth article. For each user, it will pick the article with the highest individual clickthrough probability from the remaining $N - 3$ articles and alter the ranking for $n > 3$ accordingly. Following a standard approach in the context of online news (Liu et al., 2010), the algorithm estimates click probabilities at the level of fine-grained categories by combining regularly updated data from across users (current general interest) and from within users (individual preferences). Our experiment lasts for over 20 weeks, allowing us to investigate the dynamics of economic returns to data at the individual level, and how the effectiveness of algorithmic recommendation depends on the types and diversity of data, relative to human curation. Moreover, we can also assess the dynamics behind how algorithmic recommendations *causally* influence reader choice.

Our empirical analysis proceeds in two parts. First, we are interested in the *causal* effects of algorithmic recommendation on engagement (clicks). We particularly focus on how these effects depend on the type (personal versus aggregate) and diversity (across-users) of data. In the context of across-user data diversity, we introduce the notion of *local data network effects*, where the returns to data increase with more other users, but only if the users are sufficiently homogenous. Second, we investigate whether algorithmic recommendation can change individual user consumption patterns, and which user characteristics are associated with the behavioral changes we document.

In the first part, we show that clicks on the treatment slot increase by about 4% for users in the treatment condition relative to users who see the human-curated version of the homepage. The average estimate, though, masks substantial heterogeneity along different *types of data*. When the algorithm has limited personal data, the automated recommendations based on other users' current reading behavior lead to fewer clicks than the human-curated condition. Hence, the value of algorithmic recommendations is derived from personal data as opposed to aggregate data. In particular, users at the 65^{th} percentile of visits (with 50 visits) click on the algorithmic recommendation by 18.5% more relative to human curation. While more personal data helps to increase clicks, we show that diminishing economic returns to data set in rapidly. Utilizing a coding bug, where the algorithm did not have access to updated personal data for one week, we demonstrate that it is not the stock but the constant flow of personal data that matters – though the stock can play a moderating effect. Clicks in the treatment condition become significantly negative relative to human curation during the coding bug period but less so for users who have a higher stock of prior visits. We also show that the algorithm underperforms human curation on days with surprising news developments, suggesting that human domain expertise is better able to adapt to and predict average preferences about relatively rare events.

Turning to our results on heterogeneity in data based on different user types, we show how users with different news consumption patterns react differently to the algorithmic recommendations. We base our measure of data diversity due to different user types on the average cosine distance between a user and all other users' clicks on articles across news categories over one month before the experiment. We show that clicks in the treatment group are higher for users, on average, who are closer to average consumption patterns though there is greater learning by the algorithm for users with high cosine distance. For low cosine distance users (closer to the average), we show that the algorithm performs as well as the human even if it has access only to the aggregate data and no personal data. Hence, it seems easier for the algorithm to predict individual preferences more accurately when these are less differentiated relative to the average user. Additionally, using variation in the number of users who arrive on the website on any given day, we show that this effect is more pronounced with more other users about which the algorithm can learn. However, more other users only help increase the clicks of users who are closer to the consumption patterns of the average user based on the cosine distance measure. This suggests evidence for *local data network effects*, where more data from more users only improve algorithmic performance if the users are sufficiently homogeneous.

In the second part, we utilize a measure of within-user consumption diversity based on the Herfindahl-Hirschman Index (HHI) computed on an individual user's clicks across news categories. Within-user consumption diversity is the empirical analog of an echo chamber or the concept of "going down a rabbit hole". We find that algorithmic recommendation reduces within-user consumption diversity on average. We observe interesting dynamics in this effect. As users visit the news outlet more, the consumption diversity of a user decreases significantly. This effect kicks in after a lag, which is in line with algorithmic personalization. Additionally, focusing on users who are also observed pre-experiment, an individual who had less diverse news consumption before the experiment reduced their diversity of news consumption even more. Finally, we investigate reader characteristics that might serve as moderating factors in engagement with algorithmic recommendations in the case of political news. We show that higher levels of (proxies of) digital literacy are associated with a lower tendency to reduce consumption diversity in line with popular discourse (see, for example, Susarla, 2019). Additionally, more extreme political views and lower political information levels are associated with an increased tendency to reduce consumption diversity across topics. Overall, while such algorithms increase firm revenues through increased clicks, they might prevent users from accessing diverse content, which can be detrimental from a societal point of view in the context of the news (Aral, 2020; Gentzkow, 2018).

Our study has important managerial implications and policy implications. First, in a back-of-the-envelope calculation, we convert the effects on clicks to monetary values to show that the financial returns to algorithmic recommendation can easily outweigh the implementation costs, especially for the largest news outlets. Even though diminishing returns set in with personal data, our results suggest that, depending on the type and diversity of data available to algorithms, human experts aided by algorithms can lead to higher financial returns than either of two alone. Our result on unexpected news events also points in the direction of complementarities between human expertise and algorithmic personalization. In environments where it is difficult to access personal data, managers can potentially generate a competitive advantage relative to other firms by making use of these complementary effects to offer data-driven products that are useful to consumers. Second, concerning privacy policy, there are clear trade-offs between consumer privacy and the economic value created by data (e.g., Jones and Tonetti, 2020). Our results on diminishing returns imply that a carefully calibrated personal data retention policy may provide privacy benefits without causing large reductions in consumer utility and firm revenues. Third, it is argued that data has fueled the rise of market power of online platforms, making it a central competition policy issue (e.g., Schepp and Wambach, 2016; Tucker, 2019). The argument is that data network effects (or size of user-base) can be a significant source of self-reinforcing competitive advantage (e.g., Argenton and Prüfer, 2012; Hagiu and Wright, 2020). Our results highlight that returns to algorithmic recommendations are characterized by *local data network effects*, which kick in only if users are sufficiently homogeneous.

2 Related Work and Research Framework

Our findings contribute to several streams of literature. First, we contribute to the literature investigating whether algorithms might assist or substitute human decisions. Agrawal et al. (2018) use a theoretical model to highlight that subjective judgment of humans will complement AI and ML data-driven predictions. Cowgill (2018), on the other hand, shows in the case of resume screening for labor market hires that algorithmic prediction trumps human decisions even when the outcome of interest is 'soft skills' where humans are supposed to have a comparative advantage. Using observational data and a lab experiment, Choudhury et al. (2020) show that human expertise can complement automated software tools in the context of patent examination. We add to this literature by showing that a combination of human editors' decisions, based on substantial domain expertise, and algorithmic personalization might perform the best, especially when subjective judgments are to be made, as is the case in determining the 'newsworthiness' of stories. These results further dovetail nicely with the qualitative Information Systems and Journalism literature, which analyze editorial and algorithmic judgment in news curation (Carlson, 2018; Milosavljevi and Vobi, 2019; Nechushtai and Lewis, 2019). Carlson (2018), in particular, argues that editorial judgement shouldn't be thought of as inherently subjective and suspect. Indeed, we show that editorial judgment can outperform 'objective' decisions of the algorithm when the model lacks personal data. Relatedly, there is work on algorithmic recommendations vs. curation by experts based on small scale lab settings (Senecal and Nantel, 2004; Amatriain et al., 2009; Sinha and Swearingen, 2011). Our paper is based on order(s) of magnitude larger sample size with millions of observations relative to a maximum of a few hundred (Senecal and Nantel, 2004) in such lab-based studies. This allows us to measure treatment effects very precisely which are substantially smaller relative to these studies. This is often the case with potentially small (convenience) sample studies which end up underpowered, sometimes due to a large number of treatment conditions. Moreover, we implemented our study in a randomized way in the field with individuals unaware of the existence of the experiment which can significantly mitigate external validity concerns. The estimates have direct business value and can be utilized by managers and editors at news outlets while making their decisions to hire AI engineers, editors, or implement algorithms within this context relative to studies which use outcomes such as a Lickert Scale (Amatriain et al., 2009).

Second, we complement a few empirical studies which look at the economic value of data. Chiou and Tucker (2017) analyze the effects of data retention policies, which reduced the time window search engines retain individual user data. They find that it did not affect the accuracy of search results, measured as whether the customer navigates to a new website or whether the customer repeats the search. Schaefer and Sapi (2020), on the other hand, find that the quality of search results does improve in the presence of more data on previous searches with personalized information playing a critical role. Bajari et al. (2018) analyze demand forecast accuracy in online retail and find improvements in forecast accuracy with certain types of additional data. They acknowledge the limitations of their findings by noting "... the effect that we identify may not be the true causal effect of having access to longer histories". We believe that our study is the first to provide evidence about the economic value of data based on a large scale randomized field experiment which are said to be the 'gold standard' for causal identification of key parameters (Goldfarb and Tucker, 2014). Since we observe millions of individuals repeatedly over five months, it allows us to quantify the value of different dimensions of data at the individual user level. This can help reconcile the conflicting results in this literature.[1] Moreover, we contribute *causal* evidence, from the field, to a nascent literature that discusses how measures of data quality (e.g., the freshness of data) affect how firms can create and extract value from data using theory and observational data in a descriptive manner (Brown and MacKay, 2020; Gregory et al., 2020; Valavi et al., 2020).

Next, the exercise in our paper is also related to the literature on returns to data and recommender systems in Computer Science (CS). The CS literature has evaluated the effects of different dimensions of news recommender systems (Karimi et al., 2018), typically in an offline evaluation based on historical data, which involves training data and a holdout sample (Jeunen, 2019). Offline evaluations form the basis for a large majority of research (see, e.g., the studies reviewed in Arnold et al., 2018) because such datasets are easier to acquire relative to conducting an online evaluation. Online evaluations (or randomized experiments) have gained traction because features that perform the best for recommenders offline can do very poorly when implemented in an actual randomized experiment. This is often attributed to historical data unable to account for fast-paced dynamics and customer feedback loops (Valavi et al., 2020) which is especially true in the context of online

[1] In a recent paper based on a field experiment within an online retail setting, Sun et al. (2020) also highlight that adding personal data to a recommendation engine results in higher engagement and more transactions.

news. Moreover, the CS literature has not focused on quantifying the *business value* of data though there is a related nascent literature developing related to the Shapley Value of data (see, e.g., Jia et al., 2019) with the advent of policy discussions about a “data dividend” in California (Au-Yeung, 2019). Thus, we contribute to this literature by quantifying the *causal* value of different data pieces to such systems, relative to human editors, at the individual level, which can account for dynamic feedback loops within the online evaluation. These estimates can be used in terms of their business value but also by policymakers, e.g., at the Federal Trade Commission and the European Commission, who have been concerned about access to personal data being a source of increasing returns for firms and leading to consumer privacy issues.

Fourth, our paper is also related to a literature which looks at data-driven competitive advantage at the firm- or industry level. We contribute empirical evidence to a theoretical literature in which data have positive consumption externalities and can produce a tendency towards market tipping (see, e.g., Argenton and Prüfer, 2012; Hagiu and Wright, 2020). The mechanism for such *data network effects* is typically one of across-user learning, where the firm improves the product for each customer based on what it learns from the usage of *all* its customers. Our empirical results highlight the important caveat that such learning effects depend on how homogenous the user base is. We show that adding more users leads to higher engagement only for users that show sufficiently similar historical click behavior as the average user. In relation to the concept of local network effects in geographical space or immediate social networks (Sundararajan, 2008), we introduce the novel notion of *local data network effects* in terms of the distribution of preferences or behavior.

Finally, analyzing the externalities of personalized news recommendation also contributes to the literature on the impact on diversity due to algorithmic recommendation (e.g., Oestreicher-Singer and Sundararajan, 2012; Hosanagar et al., 2014a; Lee and Hosanagar, 2018, 2019). Lee and Hosanagar (2019) finds that a collaborative filtering (CF) algorithm, which uses social data, reduced aggregate sales diversity, while *increasing* individual level diversity, within a 2-week randomized experiment on an online retail platform. Lee and Hosanagar (2018) uses the same setting to analyze the role of ‘supply-side’ or product characteristics, which moderate these CF effects. Hosanagar et al. (2014b) uses observational data to find that personalization *increased* individual-level diversity in consumption. Our study differentiates by looking at a (widely used) algorithm that combines CF and information filtering (personalization) to find a *reduction* in individual-level diversity in news consumption, which we show to be primarily due to personalization with the CF dimension having minimal impact on consumption diversity. Moreover, given that our experiment is significantly longer (20 weeks vs. 2 weeks), we can trace out how consumption diversity evolves as the algorithm sees the same reader over time. To the best of our knowledge, such dynamic algorithmic reinforcement of preferences has not been demonstrated in a causal manner previously. Additionally, due to the long time horizon of the experiment, we can analyze the moderating effects of reader- or 'demand-side' characteristics (e.g., activity on the platform, measures of digital literacy, political informedness, and political preferences), which complements the existing literature. More generally, these results, along with the dynamic reinforcement of preferences leading to a reduction in consumption diversity over time, speak to the role of algorithmic personalization in increased political polarization (e.g., Gentzkow and Shapiro, 2011; Boxell et al., 2017; Bakshy et al., 2015; Aral, 2020). Since the news is different relative to a standard retail product, dynamic reinforcement can be of great concern due to informational externalities.

3 Background and Experimental Setting

3.1 Empirical Setting: Background

Our partner news outlet is one of the largest players in the German news market with over 20 million monthly unique visitors and about 120 million monthly page impressions (total clicks) to its website. It is similar to a publication like the Wall Street Journal in size and influence and like other major news outlets, our partner gets the largest share of its revenue from advertising which makes reader engagement (e.g. clicks) crucial for its financial health. The website does not have a subscription model (paywall). In general, the German news industry seems similar in structure relative to other prominent Western democracies with a few major news outlets covering the broad political spectrum. Our partner news outlet's coverage focuses on politics, finance, and sports while also reporting on a variety of other topics. It is important to note that it is rare for major legacy news outlets in the world to experiment with algorithmic curation of their homepage. The New York Times, for instance, was among the first major news outlet to have started experimenting with personalization of an individual reader's newsfeed in June, 2019.

The decisions made by the human editor will be defined by their objective function. While it is hard to explicitly characterize this entity, it is clear from the news outlet's business model outlined above that advertising revenue will play a fundamental role. Conversations with the editors and the data science team indicate that the news outlet monitors the analytics dashboards quite closely. The fact that news editors care about increasing advertising revenue explicitly and choose news stories based on those calculations has also been highlighted extensively in the media economics literature (see Gentzkow and Shapiro, 2010, Sen and Yildirim, 2015 and discussions therein). We provide some descriptive evidence in line with this hypothesis in our data. We were able to access engagement and editorial decisions data on articles across categories. Figure 1 shows the share of all articles in a category (blue), the share of articles from a particular category making it to the homepage (red) and the share of all clicks going to articles in that category (green). It is extremely clear that the editorial decision to choose an article for the homepage is highly correlated with the amount of pageviews it garners. Hence, the divergence in the objective function between the editor and the algorithm might be limited. Hence, focusing on clicks as the main outcome variable would be a feasible way to capture various objectives in a reduced form manner.[2]

3.2 Experimental Design

The randomization procedure ensures that if a user is assigned to the control group in a particular session, then she sees the homepage curated by the human editor which involves no personalization. The layout of the website's homepage is such that the articles appear one below the other and not side by side. If the user is assigned to the treatment group, then she sees the homepage where slot 4 is personalized and the rest of the homepage remains in the same "ordinal ranking" as the control group except for this change. Figure 2 provides a simple illustration of the mechanics of algorithmic recommendation. The figure on the left shows an ordered list of 7 articles in the control group, as chosen by the human editor. On the right, the recommendation algorithm has chosen to "bump up" the article in slot 6 to slot 4. The set of articles that the algorithm can display on slot 4 comprises of the about 80 articles that are listed on the homepage at any point in time. Moreover,

[2] 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). In the extreme case, if the editor does not care about clicks at all then the estimates when comparing to a clicks driven algorithm will provide us an upper bound on the returns to data. Moreover, if there are multiple editors during a news-cycle, then we will capture the average effect across editors. We carry out a number of checks to show the robustness of our results to rule out the impact of differential editorial strategies, which is discussed in more detail below.

no algorithm is used for producing any content. Human journalists employed by the news outlet and their partners produce all the available content.

The recommendation system that was implemented uses a method developed by Google engineers for Google News (Liu et al., 2010). The objective was to show an individual user an article in their most preferred category, with the goal to increase user engagement. That is, the objective function of the algorithm is to get user i at any time t to click on the treatment slot. A user is identified based on a unique cookie ID. The system combines personal data for predicting individual interest based on past clicking behavior, and aggregate data from the clicking behavior of other users to assess current overall interests. Using past individual clicks along with current aggregate trends aimed at ensuring that while catering to individual preferences, the recommendations do not miss out on current news events. The model predicts a user's likelihood of clicking for a large number of content categories, 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 was 255, hence creating sufficient opportunity for the algorithm to identify and match a user's preferences. Each user's click history is continuously fed into the recommendation system and the prediction scores for each user and category are updated on a daily basis.[3] If the system does not have data on a user's prior clicks, then it assigns a recommendation that is soley based on current news trends as determined by aggregate data from all other users. Recommendations will increasingly be based on a users personal data as the available click history by an individual increases.[4] This feature of the algorithm allows us to distinguish between the value of personal and aggregate data. In particular, a user's engagement with algorithmic recommendations with no (or limited) prior visits to the news outlet's website will allow us to capture the value of aggregate data. As the number of visits increase, we can capture the value of personalization through accumulation of personal data. Finally, the algorithm is not (necessarily) replicating the human editor's choice for slot 4. The median recommendation was "pushed up" 10 slots.

The experiment was carried out from December 2017 to May 2018 and included all visitors, across

[3] The number of content categories available for the model overall at a point in time could be extremely fine grained and could be in the range of several hundred topics. A reader's change in preferences across topics are accounted for on a daily basis like in Liu et al. (2010).

[4] See section A of the Supplementary Appendix for more on the technical details of the implemented algorithm.

desktop and mobile devices. Similar to Aral et al. (2019) and Barach et al. (2019), the randomization is at the user-session level such that a new session commences when the user is inactive for thirty minutes and/or reloads the homepage. This level of randomization provides us with sufficient statistical power to utilize meaningful variation within users. More importantly, employing user-fixed effects allows us to control for time-invariant unobserved user-specific heterogeneity (e.g., preferences) and isolate the effect of variations in user-specific data on clickthrough rates. While this setting helps us in the clean identification of the dynamics in consumer behavior, we also use alternative levels of variation to ensure the robustness of our results in section 5.3. 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 in a paramount manner. Slot 4 gets about 3% of the total clicks on the website, but given the large overall traffic on the site and the fact that the experiment ran for multiple months, it still provides us with enough power to identify the heterogenous economic returns to data at the individual level and associated changes in consumption diversity.

4 Empirical Framework

Our baseline specification links user engagement on the website to treatment status:

$$Clicks_{is} = \alpha + \delta Treatment_{is} + \gamma_{\tau} + \mu_i + \varepsilon_{is}, \tag{1}$$

The unit of observation is user i in session s. We define a session to include all clicks that a user makes after arriving at the homepage until there is inactivity for thirty minutes or the user navigates again to the homepage. The dependent variable is either the sum of clicks that originate from the treatment slot on the homepage (*Slot=4*) or other slots on the homepage (*Slot≠4*).[5] If the algorithm performs better than the human editor, we should expect the estimate of the treatment effect δ to be positive and statistically different from zero regarding clicks to the treatment slot. The theoretical prediction for clicks on other slots is ambiguous. Even if the algorithmic recommendation outperforms the human editor on slot 4, it will depend on how attention spills over to other articles

[5] As a robustness check, we also analyze our baseline results using the logarithm of clicks, the probability of any click as well as other non-linear models in Table 6.

to determine whether there is a cannibalization or expansion effect overall. We include a day level fixed effect γ_τ to control for events affecting all users, potentially through the news cycle. We can utilize the randomization with user-level fixed effects (μ_i) and identify our effects from within-user variation. Later, we will also look at a range of specifications with alternative fixed effects and other variation to account for a variety of reading patterns as well as (human) editorial decisions to ensure the robustness of our baseline results. We cluster standard errors at the individual level to account for serial correlation of user preferences over content.

5 Baseline Results and Economic Returns to Data

5.1 Benchmark Results

We first check the validity of our randomization procedure. In Table 1, we analyze the average assignment of individuals into treatment and control groups through the experiment, based on their pre-treatment characteristics. We test the equality of means based on percentage of days active before the experiment, the total number of clicks, clicks per day, clicks during work hours and the geography of clicks across treatment and control conditions. The sample is well balanced across all observables as can be seen from the large the p-values in column (4) comparing the control (column 1) to the treatment (column 2). This indicates that our randomization has worked in the desired manner. Summary statistics of key variables are provided in Table 2, which we will use to discuss the magnitudes of our estimates below.

Next, we analyze the impact of the treatment descriptively. In column (1) of Table 3, we report the results of a simple OLS model without any fixed effects, which is equivalent to a comparison of means. We see that the number of clicks on slot 4 reduces by 1.1% with the difference between the treatment and control being statistically significant at the 1% level. In the case of column (1), we compare 0.0003 to 0.027 which is the mean number of clicks on Slot 4 in a user-session. This seems to suggest that, on average, the recommendation system cannot predict user preferences better than the human editor. However, we must exercise a bit of caution since this estimate can hide important heterogeneity. In particular, our sample includes a number of users about whom the recommendation systems has limited prior personal data and has to rely solely on aggregate data. About 20% of the overall sample accounts for only one visit by a user in the period we observe. To explore this issue further, we turn to models with fixed effects that can capture within-user variation based on the randomization.

The results in columns (2)–(4) of Table 3 paint a more nuanced picture. The specification in column (2) accounts for differences in clicking and visitation behavior across users, as well as reducing any (potential) bias in the estimates from a simple OLS model. We find that clicks that originate from slot 4 on the homepage increase by about 4% in the treatment group, compared human-curated control group. Next, we go on to explore heterogeneity in the treatment effect. The recommendation system is designed to use aggregate data on the clicking behavior of other users to recommend articles in the absence of personal data, and the algorithm gives more weight to personal data as they becomes available. This allows us to test whether the mechanism for increasing engagement is based on aggregate or personal data. We ask whether users, about whom the algorithm has more information, respond differently to the personalized recommendation by interacting the treatment dummy in equation (1) with the number of past visits, i.e. the number of times user i has visited the website since December 2017 up to that session. In results reported in column (3) of Table 3, we find that clicks to articles on the treatment slot increase with the number of prior visits, i.e. as more information becomes available to the algorithm. The average treatment effect for users without a prior visit history is negative. This indicates that the human editor is able to provide better recommendations relative to an algorithm that only has access to aggregate data. It is personal data which seems to provide the biggest payoff from algorithmic recommendation. In column (4), we look at some of the indirect effects that the experiment may have. We find that the effect on clicks to all other slots on the homepage follows a similar pattern. When the recommendation system only has access to aggregate data, the treatment effect is negative just like in the case of the treated slot. As the amount of personal data increases, so do the spillovers of clicks onto other slots on the homepage.

5.2 Economic Returns to Data

5.2.1 Characterizing the Economics Returns to Personal Data

The above results, while illustrative, are still restrictive in characterizing the returns to data since we impose that engagement responds to prior personal data in a linear fashion. We adopt a more flexible approach by running the same regression but looking at finer data bins based on the number of past visits. In particular, we run a regression of the form:

$$Clicks_{is}^{k} = \delta_1 Treat_{is} + \sum_{q} \delta_q (Treat_{is} \times PriorVisits_{qis}) + \lambda_q PriorVisits_{qis} + \mu_i + \gamma_\tau + \varepsilon_{is}, \quad (2)$$

where $PriorVisits_{qis}$ indicates whether user i in session s is in the qth percentile of $PriorVisits$ of all users across all sessions. The experimental setup at the user-session-level allows us to control for user fixed effects and the average clicks at different levels of past visits. We can therefore quantify the economic returns to data in a clean manner.

In Figure 3, we plot $\hat{\delta_1} + \hat{\delta_q} PriorVisits$ for 15 percentiles of $PriorVisits$. This provides an insightful overview of how the effectiveness of algorithmic recommendation relative to the human editor depends on personal data. The median number of visits is 16 with the 99^{th} percentile at 1189. Initially, when the algorithm has limited access to personal data, the human editor outperforms the recommendation system. That is, when there is only aggregate data available, it is optimal to defer editorial decisions to the human editor. Our results shows that up to the 35^{th} percentile of personal data, which corresponds to when the algorithm has information from up to 3 visits per user, the human editor has a comparative advantage. Around the threshold of the 40^{th} percentile (5 visits), the gap between algorithmic performance and human curation starts to become positive and economically significant.

With more available personal data, the effect of the algorithmic recommendations continues to increase. Interestingly, we see that while the effect of algorithmic recommendation increases with additional personal data, it does so at a diminishing rate. Beyond the level of the 65^{th} percentile (50 visits), click-through rates stay at similar levels of economic significance even though there might be some statistically significant differences. While diminishing returns do set in, it is important to note that an individual with 50 prior visits increases clicks by about 18.5% when they see algorithmic recommendations relative to human curation. This is an economically meaningful number, which can have significant revenue consequences for the news outlet, an issue which we delve into deeper below.

Overall, these results suggest that it is personal data and not aggregate data that generates the most value. That is, the biggest payoffs in algorithmic aided decision-making might come from *per-*

sonalization as opposed to simple *automation* based on current overall trends. It is also insightful to see that the returns to data results in a smooth curve without any obvious discontinuities, threshold effects or step functions which, as highlighted above, has been of concern from a policy perspective. Moreover, this figure shows that personal data can help firms gain a competitive advantage but diminishing economic returns do set in quickly. Hence, the competitive advantage for firms from employing data in algorithms might be limited beyond a point. We will expand on this issue in the next section. On the flip side, this has interesting implications for privacy policy. Regulatory efforts, like the European General Data Protection Regulation or the California Consumer Privacy Act, that can restrict the amount of personal data that firms can collect and process, might not erode much data-driven value for firms and consumers, since adverse consequences on consumer engagement would be limited. Additionally, since we can estimate the value of data based on the number of visits to the website, we can relate our general characterization to the prior literature. For example, Chiou and Tucker (2017) find no change in search engine precision, measured by click through rates, when European regulation forced search engines to retain individual level data for a much shorter period of time (Yahoo!-13 to 3 months while Bing-18 to 6 months). This result can be rationalized by realizing that the gains from data might have been on the 'flat' part of the curve in Figure 3 where additional data does not add much to the overall effect.

5.2.2 Data Diversity and Local Data Network Effects

The results above highlight the important role of personal data. The returns to data curve shows how the effect of the algorithm depends on the amount of personal data for the average user. Next, we analyze the impact of user types or heterogeneity in reading behavior and its impact on data availability for algorithms. For a general news website, there can be significant heterogeneity in reading behavior across users which can impact the learning and targeting ability of the algorithm. An analogous argument has been made in the case of privacy regulation and cookie based advertising by Goldfarb and Tucker (2011), when comparing websites with specific and general content. We investigate how the returns to aggregate and personal data depend on across-user heterogeneity, i.e. the diversity of data available to the algorithm. We use a standard measure of heterogeneity in user preferences: cosine distance. To make sure that our measure of across-user heterogeneity is not affected by the treatment, we focus on the set of users who we observe before the experiment. We define the average cosine distance between a user and all other users clicks on news articles across news categories over one month before the experiment. A user with a high cosine distance has a more diverse reading behavior relative to the average user.

Using this measure, we estimate the returns to data conditional on across-user heterogeneity by augmenting equation (2) such that

$$\begin{aligned} Clicks_{is}^{k} = {} & Treat_{is}\left(\delta_1 + \sum_q \delta_q PriorVisits_{qis} + \sum_q \theta_q (PriorVisits_{qis} \times Dist_i)\right) \\ & + \lambda_q PriorVisits_{qis} + \gamma_\tau + \mu_i + \varepsilon_{is}, \end{aligned} \tag{3}$$

where the base effects of $Dist_i$ and $PriorVisits_{qis} \times Dist_i$ are absorbed by the user fixed effect μ_i. As before, we plot $\hat{\delta_1} + \hat{\delta_q} PriorVisits + \hat{\theta_q} PriorVisits \times Dist$ for 15 percentiles of $PriorVisits$ in Figure 4, but for the 10th and 90th percentile of the pre-experiment cosine distance in the sample. First, for users with a low average cosine distance (10th percentile), at low levels of personal data, the effect of algorithmic recommendations is not different relative to human curation. However, we do see a negative treatment effect for users with a high cosine distance (90th percentile). This implies that the negative overall treatment effect for low levels of personal data, seen in Figure 3, can be explained by heterogeneity in reading behavior. When the recommendation system has little personal data and relies more on aggregate data, user heterogeneity plays a key role in how intensively users respond to recommendations. Second, we find that the treatment effect is consistently larger for users with low cosine distance than it is for users with high cosine distance. Hence, this seems to confirm that it is easier for the recommendation system to pick up on more mainstream preferences than on niche preferences. At the same time, the effect varies significantly more with additional data for users with high cosine distance (from -0.01 to 0.005) than for users with low cosine distance (0 to 0.0075), suggesting that there is more algorithmic learning for users with niche preferences. This, in fact, is confirmed by the regression results in columns (5) and (6) of Table 3 which captures the average effects of estimates in Figure 4. Focusing on the same sample of users who are observed in both the treatment and the pre-treatment period. We can see in column (5) that individuals with greater heterogeneity in reading behavior have a significantly lower treatment effect since the coefficient of Treatment × Cosine Distance is negative and statistically significant. Additionally, in column (6) the coefficient of Treatment × Cosine Distance × Prior Visits is positive indicating that there is more learning by the algorithm as it observes individuals with greater reading heterogeneity repeatedly.

Next, we can look at our experimental setup through the lense of data network effects (across-user learning). The idea of data network effects is that a larger user base provides more data about product usage, from which the firm can increasingly learn and improve product quality, which attracts additional users. Hence across-user learning will lead to a re-inforcing competitive advantage and ultimately market tipping (see, e.g. Argenton and Prüfer, 2012; Hagiu and Wright, 2020). In our context, we can test whether returns to algorithmic recommendation are higher with a larger user base. Of course, the effect of a larger user base will show up strongest when algorithmic recommendations are still based on other users' data (aggregate data), and much less so when based on personal data. In the context of our experimental setup, we focus on the first 6 visits of a user (40^{th} percentile of visits) to capture the effect of algorithmic recommendations when aggregate data still play a role and such that we have sufficient statistical power to pin down estimates precisely.[6] We then explore variance in the number of users that comprise the aggregate data that the algorithm can access. We exploit daily variation in the number of unique visitors to see how the clickthrough rate of users in the treatment group differs conditional on the number of users that visited the website on the previous day. The model specification is similar to equation (3) – we estimate the conditional treatment effect for users with low and high levels of distance to the average user. Figure 5 shows some clear patterns related to potential data network effects. As the (lagged) number of users on the website increases, engagement with algorithmic recommendations based on aggregate data increases only for individuals with a low cosine distance. This implies that as the user base expands, only individuals whose reading behavior is sufficiently close to the average preferences click more. Quantitatively, this means that a 1% increase in size of user base leads to a 6.6% increase in the engagement with algorithmic recommendations. For individuals with high cosine distance, the change in the number of users leads to essentially no change in the treatment effect. Even though these are estimates based on short-run fluctation in the user base, there is sufficient variation since this ranges from about 430,000 to 750,000 conditional on the general news

[6] This result is qualitatively robust to alternative thresholds. Increasing (decreasing) the number of visits used in the estimation makes the curves take similar shapes and tightens (widens) the confidence bands by providing more (less) power. These results are available upon request.

interest on a given day.

We conclude that this suggests the presence of *local* data network effects. That is, the size of the user base may provide a strategic advantage *only if* the new users have sufficiently homogeneous preferences compared to existing users. This provides a caveat to the recent theoretical literature, in which user preferences are typically assumed to be homogenous. At least in our setting, user heterogeneity seem to be a key empirical determinant of the strength of data network effects. We find that only users with the lowest levels of cosine distance react stronger to algorithmic recommendations that are based on other users's data when there are more users whose data the recommendation system can access.

5.2.3 Stock vs. Flow of Data using a Natural Experiment

In this section, we analyze an additional facet of the value of data, which is the time dimension. A key characteristic of the news industry is the high frequency in production (number and types of stories) and demand (reading within and across news categories). Each visit to a news website creates a *flow of data*, which over time accumulates to a *stock of data* on click behavior. The dimension of stock vs. flow is of strategic importance, because if the stock of data is sufficient to predict user preferences well enough, then an incumbent can have a significant advantage over an entrant. On the other hand, if the flow of data is consequential, then entry barriers are lower. The role of stock vs. flow in creating value from data might not only be important in news, but also in other high frequency industries such as finance, search and online advertising, as well as for platforms with user-generated content or connected device data.

To analyze the stock vs. flow dimension in our context, we tap into an additional source of exogenous variation by exploiting a natural experiment within our field experiment. In particular, we make use of a coding bug in the implementation of the recommendation system at our partner news outlet that remained unnoticed for 6 days in January 2018. The bug impaired the ability of the algorithm to access daily updates of personal data. The coding error was such that 2017 was hard-coded as the year of the historical data to be used to make user-specific predictions. Since there was no personal data available for January 2017, no fresh personal data was loaded and recommendations in the first week of January 2018 were based on the latest available personal data from December 2017. As time went on, the personal data utilized by the algorithm became increasingly outdated. If it is true that 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 1*st* of January would see more ‘relevant’ content than if they visited on the 6*th* of January. If the existing stock of data is sufficient to predict individual preferences well enough, however, we should not see that users engage less with the algorithmic recommendations. The human editor was of course not affected by the coding bug. Since the control group continued to see the human curated version of the homepage, the coding bug allows us to isolate the effect of less recent personal data from other confounding factors, such as the news cycle in the first week of January.

We focus on data from December 2017 and January 2018 of our sample. We define the indicator *New Year Bug* equal to one on January 1–6, 2018 when the bug in the code went unnoticed or zero otherwise. We will focus our attention on the interaction term of whether a user was in the treatment group and whether the user was observed during the *New Year Bug* period. Column (1) of Table 4 shows that clicks on slot 4 are significantly lower when a user is in the treatment relative to the control during New Year Bug period. The total effect even turns negative and corresponds to a reduction of 11%. In column (2), instead of the New Year Bug indicator, we introduce a linear trend capturing each successive day that went by with the bug remaining unnoticed. We find that this interaction term is significantly negative as well, which implies that with an additional day without updated data, users in the treatment group clicked on algorithmic recommendations significantly less. Finally, we assess whether the stock of existing data can mitigate some of this reduced precision in personalization. We create a variable which captures the number of visits an individual had before the January 1st 2018 as the stock and analyze its interaction with the treatment focusing on the six days in which the bug was left undiscovered. The result in column (3) shows that the interaction effect is positive which means the stock can help predict user level preferences and mitigate, to some extent, the need for continuous updating.

Overall, this shows that it is not only the stock but the constant flow of data which also matters in generating engagement with algorithmic recommendations. While the stock of personal data can moderate these effects, this exercise highlights how demand-side behavior varies in a dynamic environment. Finally, since the exogeneous variation induced by the New Year Bug only affects personal data, this exercise serves as additional evidence that our baseline results are indeed reflecting the returns to personal data.

5.2.4 Surprising Developments, Editorial Judgment and Algorithmic Predictions

In the results above, the human editor gains a competitive advantage over the algorithm because of limited personal data. In this part, we attempt to test a similar idea in the case of significant 'surprising' news event days. Due to limited data on such news events, it can be envisaged that human editors, who have domain knowledge and expertise, are better at forecasting the 'newsworthiness' of a potentially fast paced, developing story. More generally, it is important to note that being 'first to market' with a surprising news development is a crucial source of revenue for media outlets (Franceschelli, 2011). Additionally, for a significant proportion of stories, competing outlets catch up with the news outlet reporting the facts on the the big story in a matter of minutes (Cagé et al., 2020). Hence, this is an important dimension along which news outlets would need rapid precision. We explore this dimension by analyzing 'surprising' developments related to the formation of the coalition between parties after the German federal elections in early 2018. We further investigate sudden spikes in public interest in sports, such as gold medals for German athletes in the Winter Olympics 2018. Figure 6 illustrates these spikes in public interest on particular days with respect to politics and sports in Google search data. In Table 5, we provide evidence in favor of the idea that human judgement beats the algorithm in specific cases. The results in columns (1) and (2) show that clicks to politics articles on the treatment slot decrease for users in the treatment group on dates with surprising news events. We repeat this exercise for surprising sport events and clicks to sports articles on the treatment slot to find very similar results in column (3). These results are robust to alternative time thresholds for these events.

We can also show that these results hold more generally. We construct a demand-side measure of 'surprising' news days. 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. This gives us a measure of the demand variation of each category on a given day. Based on this, we create two variables that let us characterize how 'surprising' 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 higher than 4 (i.e., higher than the 95th percentile). This is true for 20%

of the days in our sample. Second, we define the continuous variable *DemandVariance* to be equal to 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 1.16). The 5th percentile is 1.74, whereas the 95th percentile is 5.06. In columns (4) and (5), we look at clicks to the treatment slot over the entire sample, and show that the treatment effect is significantly smaller on high variance days. This suggests that the editor is better able to judge the inherent 'newsworthiness' of a particular day. Hence, whatever information an editor is using for this judgement is either unobserved or not explicitely used in the algorithm in our field experiment.

This interpretation is in line with related work on complementarities between humans and ML systems. For example, Choudhury et al. (2020) analyze the performance of a simple ML technology that helps patent examiners to locate prior art. They conclude that "individuals who possess domain expertise [...] are complementary to ML in mitigating bias stemming from input incompleteness" (Choudhury et al., 2020, p. 3).

5.3 Validation, Alternative Variation and Robustness

5.3.1 Alternative Sources of Variation and Robustness

While the estimates in the previous section are based on fine-grained data from a randomized experiment, a few issues might impact the results.

First, we run alternative econometric models in Table 6. In column (1) we change the dependent variable to a logarithmic transformation of clicks to slot 4 ($\ln(1 + Clicks_{is}^{T})$). In column (2) we estimate a linear probability model where the dependent variable indicates at least one click to slot 4. Both approaches yield qualitatively similar results to the baseline. In columns (3) and (4), we use a Poisson model and a negative binomial model to find a similar positive and statistically significant average treatment effect.

Second, we carry out a few checks of the baseline specification using alternative fixed effects, reported in the appendix, to ensure the robustness of our results. While we have user- and day-fixed effects separately, this might not account for different reading behavior at different times of the day, or the fact that different human editors might operate at different times. Columns (1) and (2) in Table 7 in the appendix show that results using within user-week and user-day variation are qualitatively and quantitatively similar to our baseline estimates. Controlling for user-hour and user-hour-of-the-day variation in column (3) and (4) yields remarkably similar results. This clearly suggests that our baseline results are not driven by differences in reading behavior due to changes in the news cycle or by different preferences or strategies adopted by human editors.

Third, in columns (1)-(3) of Table 8, we use some more sources of variation to further assess the robustness of our results. In column (1) we restrict the data to look at only the first session of every hour for a user to find qualitatively similar results. In column (3), we use the first session of the morning for a user (6 am to 12 pm) while column (4) uses the first session of the afternoon (12pm to 6pm) to find similar results as in the baseline. For each of these specifications in (1)-(3), we find that the results are qualitatively and quantitatively in line with our baseline estimates. This gives us confidence that our results are driven by the treatment and do not pick up demand dynamics.

5.3.2 Rank Effects and Personalization

Throughout the analysis we have provided different pieces of evidence to show that personalization provides the biggest payoff from algorithmic recommendations. The underlying idea is that the algorithm moves up articles which the human editor has placed further down the homepage. Moreover, this process works better the more personal data the algorithm has since it can find a better fit with individual preferences. We attempt to get at this by looking at snapshots of data of the website's homepage layout at five minute intervals for a number of days during the experimental period. A caveat here is that the analysis will be correlational since we can only exploit the variation within the treatment group over time. Utilizing this data, we show that results in columns (4) and (5) of Table 8 are in line with intuition. In column (4), we can see that, on average, an article that has a lower rank in the control condition (a higher number) will lead to less clicks when moved to slot 4 in the treatment condition than other articles on slot 4. This is in line with the editor placing an article lower down 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 which are ranked 14 or lower (at least 10 ranks lower than the treated slot) to ensure the "pulling up" of articles will be meaningful in that the readers might not have seen the article otherwise. The results are qualitatively similar implying that an article that would not have been noticed before is noticed when ranked higher and combined with better fit, leads to more clicks on the treated slot. Overall, this exercise does provide more evidence in line with the proposed personalization mechanism.

5.3.3 Additional Evidence from a Randomized Online Survey

Our field experiment was set up such that users were not informed that they receive personalized recommendations. It may still be the case that some users have realized that they were shown automated recommendations. A concern could be that these users clicked on the treatment slot simply because they liked that they were treated with a customized experience, and not because they liked the customized content.

To rule out this alternative explanation, we carry out an online lab experiment. We mimic the looks of our main field experiment, but change the treatment such that we simply label slot 4 as “Recommended for you”, but leave the content the same for the treatment and control groups. We list a total of 6 news articles with headline and teaser text, exactly as they appeared on our partner news outlet. When we carried out the experiment in May 2020, the news were dominated by the COVID-19 pandemic, occupying most top slots on the homepage of our partner outlet. We therefore use news stories from May 2019 and tell participants to envision being back in 2019 in a pre-study vignette.

We recruit 1,500 German internet users through the crowdsourcing platform Clickworker – the German pendant to MTurk. In a first step, we ask participants to choose their favorite news categories. We don’t use these but to make our “recommendations” more credible in the subsequent step. A user can be assigned to a “curated” condition where we highlight that the slot 4 story was chosen by the Editor. In another condition which reflects personalization, we label the slot 4 story as “Recommended for you”. Finally, we have a control condition with no label. In all of these conditions, we serve the participants an article which is randomly chosen and not tailored to their stated preferences. The idea is to see whether a simple ‘label’ can generate a differential click behavior without the article on slot 4 actually matching the participant’s preferences. We repeat this exercise three times for each participant, using six different news articles each, where the participants are randomized into treatment and control conditions in every repetition. We can therefore use user-fixed effects to control for unobserved preference heterogeneity. Our final sample has 1,413 users. The results in Table A.3 in the Appendix, show that we find null effects for each treatment condition on either the treated slot (column 1) or the untreated slots (column 2). This is not driven by a lack of power, since pooling all slots together in column (3) leaves the null results unchanged. Overall, this suggests that our result of algorithmic recommendation matching reader preferences cannot be generated mechanically by simply labeling (or not).

5.4 Revenue Implications for Automating Editorial Curation

Our analysis provides a coherent picture of the extent to which an algorithm can outperform the human editor and how this might crucially vary with the amount of data available. In this section, we try to put our estimates' economic size into context with a simple back of the envelope calculation. In particular, we want to assess how much revenue a news outlet might generate if they completely automated the curation process, relative to the costs of implementing such a system, including hiring data scientists.

We make the simplifying assumption that if the news outlet automates the entire curation process, then the increase in overall clicks will be the same percentage as observed in the experiment. This magnitude is 3.75%, which we take from column (2) of Table 3. This news outlet gets about 120 million clicks per month, which means that a 3.75% increase in clicks will lead to an additional 4.5 million clicks every month. The average click-through rate on display is about 0.35%, which implies total additional monthly clicks on a particular ad would be 15,750, and with two prominent ads on each page, the total clicks on ads would be 31,500.[7] The average cost per click is about $0.58, which implies that the total additional monthly revenue accruing to the news outlet from automating editorial curation is $18,270. The average monthly salary of a data scientist in Germany, along with benefits, comes to a total of $7,200 (approx.) which can be considered as the cost of implementing such an automated system. Even after accounting for this potential cost, our estimates suggest that it would be profitable for the news outlet even if the data scientist works on this task full-time. To provide some more context, a news outlet with aggregate traffic of about 47.5 million clicks per month will be able to break even, which corresponds to the median of the top 50 news outlets in Germany. In other words, the top 25 news outlets in Germany will find such automation

[7] 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.

profitable.[8] To make some comparisons with international outlets, monthly clicks to the Wall Street Journal, Los Angeles Times, and Boston Globe are around 120 million, 63 million, and 30 million, respectively.[9]

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

6 Information Externalities in Algorithmic Recommendations

6.1 Consumption Diversity and Personalization

News is a special product because of its public good nature. In particular, since the treatment effect is driven by personalization based on the recommendation algorithm being trained on prior individual data, which is "biased" towards personal preferences and hence, could be at odds with "socially optimal" reading behavior.[10] The consumption of some types of articles could be deemed more socially valuable because it may lead to better informed political decisions (e.g., voting) of individuals. Hence a shift in the distribution of readership across article types can have welfare implications beyond the firm's intentions. We will analyze how algorithmic recommendations might have affected browsing behavior across different types or categories of articles over the experimental period. We are interested in the impact of the clicks on the algorithmically recommended article, which would be the result of a spillover from the recommendation on the overall consumption diversity. Hence, we aim to quantify the change in consumption diversity across broad article categories (such as politics, sports, finance, etc.) coming from clicks on slot 4 and the resulting spillovers onto other slots.

A priori, the direction of change in consumption diversity, if any, is ambiguous since the algorithm is trained on past individual reading behavior and aggregate news trends based on the browsing

[8] See http://ivw.eu/englische-version for details on these numbers.

[9] See https://www.similarweb.com/website/bostonglobe.com for aggregate statistics on this. Additionally, 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.

[10] Of course, it is hard to define what "socially optimal" is, but in popular discourse, it often ranges from 'hard' vs. 'soft' news as well as 'partisan' vs. 'objective' news. These terms come into play in the mainstream media because of the importance of information externalities through the news.

behavior of other readers. Hence, there could be increased or decreased consumption diversity based on a reader's initial preferences relative to the rest of the users, the rate of change in individual reading behavior, and changes in the supply of stories through the news cycle. The baseline results described above, though, might imply that there is a reduction in individual-level consumption diversity since the treatment effect is driven primarily by personalization.

We use the Hirschman-Herfindahl Index (HHI) measure of consumption shares across different broad categories at the individual level. HHI is a commonly used measure of market concentration in the Industrial Organization literature. In a standard setting with firms, the HHI is the sum of squares of market shares across firms, where market shares are defined as fractions. It considers the relative size distribution of the firms in a market, and hence, if the market is controlled by one firm, then the HHI will be equal to 10,000. The HHI will approach zero if the market has a large number of equal-sized firms. The HHI will increase if the number of firms in a market decrease or the disparity in size between a given number of firms increases. We map this definition into our context of reading "concentration" across categories. Since the article categories remain unchanged through our sample period, an increase in HHI would result from an increase in the disparity of the relative distribution of clicks across categories from a reader, implying an increase in concentration across categories.

Since our randomization takes place at the user-session level, we create two observations per user, which calculates the HHI whenever the user was in the treatment and control group separately throughout the sample period. We then regress these HHI measures on the treatment variable to assess how browsing behavior differed on average across all users. The results in Table 9 show that the HHI increased when the users were in the treatment group relative to the control, suggesting that the recommendation algorithm leads users to find similar categories to those recommended. The magnitudes imply that there was an increase in user-level HHI by 6% for slot 4. This result is in line with personalization catering to the individual's preferences and hence, reducing diversity in consumption. These magnitudes could be a cause for concern in the traditional sense given that an increase in HHI by 200 points in a 'highly concentrated' industry is considered problematic. The mean HHI measures in our sample are about 7500, which is high. This discussion, though, is to give the reader a better context for the magnitudes.

Next, we use pre-experimental browsing behavior from November 2017 for individuals we observe before the experiment to assess how their consumption diversity is affected by personalized recommendations. In column (2), we show that individuals who had less diverse consumption in the pre-experiment period narrowed their consumption diversity even further during the experiment when in the treatment group. This can be seen by the positive interaction between the treatment and the pre-user-HHI. Finally, in column (3), we delve into a potential mechanism by interacting the treatment with the number of visits to the website. As can be seen, as individuals visit the website more often, their consumption diversity declines further. Again, this result is in line with the increasing personalization as the user comes back to the website repeatedly.

Next, we attempt to dig deeper into the mechanism and dynamics that could drive reduced consumption diversity. In particular, we exploit the relatively long time dimension of our experiment and expand on the result of column (3) in Table 9. In particular, we analyze how individual-level HHI varies based on a user's visits to the website.

Figure 7 plots the average user-level HHI on the treatment slot using variation in visits across users in the treatment relative to the control group. There are two takeaways from this figure. Initially, when recommendations are based on aggregate social data, there is a limited impact on individual consumption diversity. As the impact of personalization kicks in, with an increase in visits, we see a rise in HHI, implying a reduction in consumption diversity. Moreover, this increase in HHI appears to continue, albeit at a diminishing rate, even at higher levels of visits. This result on the dynamics could be of importance since we demonstrate causally how algorithmic recommendations drive an individual into a "rabbit hole" by continuously reinforcing their preferences over time. Such dynamics have been hypothesized about (Aral, 2020), and our paper is the first to provide causal evidence. As mentioned before, this can have significant consequences, especially in the case of political news.

6.2 Political Content and User Characteristics

Finally, we assess the characteristics of readers who are more prone to "go down the rabbit hole" and reduce consumption diversity due to recommendations. Such a tendency has often been attributed to a lack of digital literacy with the new "digital divide" being an "algorithmic divide" (Susarla, 2019). Individuals with extreme political views and a lack of political information are also associated with such behavior.[11] Analyzing these heterogeneous treatment effects can be an informative exercise to provide evidence for the public debate. We test for these hypotheses by using proxies for such characteristics. Since our data does not include individual-level covariates that let us directly classify a user's digital literacy level, we conduct a supplementary survey. We access a panel of 500 German internet users through the crowdsourcing platform Clickworker. Looking at the 497 usable observations, we conclude that our respondents are very similar in age, education, and income compared to internet users in the German Socio-Economic Panel (SOEP), which is well known as representative of the German population (Wagner et al., 2007). The average age of an internet user in SOEP is 39, in our data 37. In both data sets, the average internet user has completed secondary education, and the average personal net monthly income is between 1,500 and 2,000 EUR. We construct an index of digital literacy using five survey questions (see Table A.1). We further ask participants whether they read news online and which device they use to do so (smartphone, tablet, laptop/desktop). The simple OLS model in Table A.2 shows that not using a laptop/desktop to read news online is a strong predictor of lower levels of digital literacy, even after controlling for age, education, and income. We use this information as an individual-level proxy for digital literacy. Relatedly, Qian et al. (2019) find that individuals using only mobile searches for online shopping are based in more economically dis-advantaged regions.

Results in column (1) of Table 10 suggest that individuals that did not access the news website through a desktop or laptop computer are more likely to increase their consumption share of politics when treated. We use state-level voting as proxies of extreme political views by looking at the share of votes going to the extreme right and left parties in the federal election of 2017. Following Larcinese (2007), we use voter turnout in the last election as a proxy for political knowledge, aggregated at the state level. Results in column (2) of Table 10 suggest that individuals who reside in states with a high share of votes to extreme parties are more likely to increase their share of clicks on political stories when in the treatment. Additionally, regions with a higher voter turnout, a proxy for being more informed, are more likely to increase their click share for political news as seen in column (3).

Overall, we aim to provide some grounding for assertions put forward in the public discourse with these results.

[11] See, e.g. https://tinyurl.com/yybl4n58.

7 Discussion and Conclusion

We run a field experiment to quantify the economic returns to data and informational externalities associated with algorithmic recommendation relative to human curation in the context of online news. We partner with a major news organization in Germany to implement a recommendation algorithm that is based on a combinaton of aggregate information about consumer preferences and personal data. Our baseline results show that algorithmic recommendation can outperform human curation in terms of user engagement. We find significant heterogeneity with regards to the *type of data* as personal data is much more effective than aggregate data, but diminishing returns set in quickly. Despite these diminishing returns, we find economically significant effects. For example, for users with personal data at the 65^{th} percentile, algorithmic recommendations lead to a 18.5% increase in clicks. Additionally, our results show that *data freshness* is crucial. We find that a lack of updating of personal data quickly detoriates users' engagement with algorithmic recommendations. We can also show that the algorithm underperforms human curation on days with surprising news developments. All these results suggest that during a time where there is a lot of discussion about which tasks will be automated in the newsroom, algorithms can complement human editorial expertise. This complementarity might play out especially when humans are better able to adapt the decision-making process in fast-paced and rapidly changing environments with dynamic feedback. Such fast-paced environments are increasingly prevalent in the context of online platforms in part, due to the use of algorithms (Brown and MacKay, 2020). In a discussion of managerial implications, we further contrast these benefits with the costs associated with the implementation of an algorithmic recommendation system. Our findings also contribute to the recent policy debate related to privacy concerns and data-driven competitive advantage, particularly of large firms. In particular, if data retention is to be limited due to privacy concerns, then our results suggest that a carefully calibrated policy will not hurt the economic effectiveness of algorithmic recommendation much due to potentially diminishing economic returns to personal data.

We highlight important heterogeneity in the returns to data. We find that users closer to average consumption patterns engage more with algorithmic recommendations. We further provide evidence for the existence of *local data network effects* where individual engagement only increases with the size of the user base for users close to average consumption patterns. The size of the user base is said to be of great value for firms to create data-driven services and products. Our results suggest that with sufficient variation in incoming data, e.g. from diverse preferences of consumers, firms might not benefit as much from externalities created by additional users. This provides a significant nuance to results in the theoretical literature, relevant for managers and policy makers alike. Data network effects may only be a driver of market power of online platforms, which a central concern for competition policy, in cases where the user base is sufficiently homogenous.

Finally, we find that algorithmic recommendation, which is mostly based on personal data, leads to a reduction in the diversity of news consumption. This also relates to other contexts and is in stark contrast to Hosanagar et al. (2014a) who look at collaborative filtering recommendations (i.e. based on data from other users) in online shopping behavior and find that individual-level diversity increases. The context of consumption diversity in the news is important in its own right due to the concern related to "filter bubbles" and resulting informational externalities in an increasingly polarized society. Additionally, we are able to characterize the dynamics of this reduction in diversity of news consumption which has been central to the public debate but with limited empirical evidence. We also show that proxies of low digital literacy and extreme political views are associated with higher engagement with algorithmic recommendation. We believe that these results are important in demonstrating behavioral patterns which are at the core of a recent debate (Aral, 2020; Gentzkow, 2018).

We conclude by highlighting some limitations of this paper leading to avenues which should create opportunities for future work. First, as with any field experiment, our results are based on one experiment we carried out with one partner firm using one algorithm. While we believe that our results could apply in a variety of different settings, questions of external validity would always be pertinent. Given the significance of this research area both for firm strategy and public policy, we need to create a menu of evidence related to the value of data. A particular dimension to analyze would be to look at alternative algorithms or broaden the range of decisions the algorithm was allowed to take. In our setting, the algorithm had a relatively simple curation task based on a set of articles already chosen by the human editor for the homepage, creating a natural combination of human decisions and ML. It would be interesting to see how the returns to data would be different if the algorithm could have access to a broader set of articles to choose from or could even create its own articles. With the rise of 'robo-journalism', the ability of algorithms to write reasonably sophisticated articles is making its way into the newsroom slowly but surely. Next, a dimension of this study that could be expanded upon is to look at firms which have multiple, different types of products which they sell to different types of consumers across different platforms. It would be informative to understand how additional data on an individual user across products affects algorithmic predictions. This would also bring us one step closer to understanding concerns associated with potential increasing returns to data accruing to large platforms such as Google and Facebook. Finally, it would be important to better understand which types of consumers are likely to engage more with algorithms and as well as the dynamics of reinforcement. Our analysis brings out some interesting patterns but is the first step in understanding this phenomenon. This issue is becoming increasingly important with the use of such recommendation systems by companies such as Youtube, who are being accused of pushing more "extreme content" or "fake news" to garner user engagement.

Notwithstanding these limitations, we believe that using an algorithm from a well-cited paper within a field experiment in a large news outlet does contribute to our understanding of editorial judgement, the value of data in algorithmic recommendations and its impact on consumption diversity.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The simple economics of artificial intelligence*. Harvard Business Press.

Amatriain, X., Lathia, N., Pujol, J. M., Kwak, H., and Oliver, N. (2009). "The wisdom of the few: a collaborative filtering approach based on expert opinions from the web." In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 532–539.

Aral, S. (2020). *The hype machine: how social media disrupts our elections, our economy, and our health–and how we must adapt*. Currency.

Aral, S., Eckles, D., and Kumar, M. (2019). "Scalable bundling via dense product embeddings." *MIT Working Paper*.

Argenton, C., and Prüfer, J. (2012). "Search engine competition with network externalities." *Journal of Competition Law and Economics*, *8*(1), 73–105.

Arnold, R., Marcus, J. S., Petropoulos, G., and Schneider, A. (2018). "Is data the new oil? diminishing returns to scale." *Working Paper*.

Au-Yeung, A. (2019). "California Wants To Copy Alaska And Pay People A Data Dividend. Is It Realistic?" *Forbes*, *https://www.forbes.com/sites/angelauyeung/2019/02/14/california-wants-to-copy-alaska-and-pay-people-a-data-dividend--is-it-realistic*.

Bajari, P., Chernozhukov, V., Hortaçsu, A., and Suzuki, J. (2018). "The impact of big data on firm performance: An empirical investigation." *Working Paper*.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). "Exposure to ideologically diverse news and opinion on facebook." *Science*, *348*(6239), 1130–1132.

Barach, M. A., Golden, J. M., and Horton, J. J. (2019). "Steering in online markets: the role of platform incentives and credibility." Tech. rep., National Bureau of Economic Research.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2017). "Greater Internet use is not associated with faster growth in political polarization among US demographic groups." *Proceedings of the National Academy of Sciences of the United States of America*, *19*, 1–6.

Brown, Z., and MacKay, A. (2020). "Competition in pricing algorithms." *Available at SSRN 3485024*.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). "The production of information in an online world." *The Review of Economic Studies*, *87*(5), 2126–2164.

Carlson, M. (2018). "Automating judgment? algorithmic judgment, news knowledge, and journalistic professionalism." *New media & society*, *20*(5), 1755–1772.

Chiou, L., and Tucker, C. (2017). "Search engines and data retention: Implications for privacy and antitrust." *Working Paper*.

Choudhury, P., Starr, E., and Agarwal, R. (2020). "Machine learning and human capital complementarities: Experimental evidence on bias mitigation." *Strategic Management Journal*.

Cowgill, B. (2018). “Bias and productivity in humans and algorithms: Theory and evidence from resume screening.” *Working Paper*.

Franceschelli, I. (2011). “When the ink is gone: The impact of the internet on news coverage.” Tech. rep.

Gentzkow, M. (2018). “Media and artificial intelligence.” *Working Paper*.

Gentzkow, M., and Shapiro, J. M. (2010). “What drives media slant? evidence from us daily newspapers.” *Econometrica*, *78*(1), 35–71.

Gentzkow, M., and Shapiro, J. M. (2011). “Ideological Segregation Online and Offline.” *Quartely Journal of Economics*, *126*(4), 1799–1839.

Goldfarb, A., and Tucker, C. (2011). “Online display advertising: Targeting and obtrusiveness.” *Marketing Science*, *30*(3), 389–404.

Goldfarb, A., and Tucker, C. (2019). “Digital economics.” *Journal of Economic Literature*, *57*(1), 3–43.

Goldfarb, A., and Tucker, C. E. (2014). “Conducting research with quasi-experiments: A guide for marketers.” *Rotman School of Management Working Paper*, (2420920).

Gregory, R. W., Henfridsson, O., Kaganer, E., and Kyriakou, H. (2020). “The role of artificial intelligence and data network effects for creating user value.” *Academy of Management Review*, (ja).

Hagiu, A., and Wright, J. (2020). “Data-enabled learning, network effects and competitive advantage.” *Working Paper*.

Hosanagar, K., Fleder, D., Lee, D., and Buja, A. (2014a). “Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation.” *Management Science*, *60*(4), 805–823.

Hosanagar, K., Fleder, D., Lee, D., and Buja, A. (2014b). “Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation.” *Management Science*, *60*(4), 805–823.

Isaac, M. (2019). “In new facebook effort, humans will help curate your news stories.” *New York Times*, https://www.nytimes.com/2019/08/20/technology/facebook--news--humans.html.

Jeunen, O. (2019). “Revisiting offline evaluation for implicit-feedback recommender systems.” In *Proceedings of the 13th ACM Conference on Recommender Systems*, 596–600.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. (2019). “Towards efficient data valuation based on the shapley value.” In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176.

Jones, C. I., and Tonetti, C. (2020). “Nonrivalry and the economics of data.” *American Economic Review*, *110*(9), 2819–58.

Karimi, M., Jannach, D., and Jugovac, M. (2018). “News recommender systems–survey and roads ahead.” *Information Processing & Management*, *54*(6), 1203–1227.

Larcinese, V. (2007). “Does political knowledge increase turnout? evidence from the 1997 british general election.” *Public Choice*, *131*(3-4), 387–411.

Lee, D., and Hosanagar, K. (2019). “How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment.” *Information Systems Research*, *30*(1), 239–259.

Lee, D. D., and Hosanagar, K. (2018). “How do product attributes and reviews moderate the impact of recommender systems through purchase stages?” *Kartik, How Do Product Attributes and Reviews Moderate the Impact of Recommender Systems Through Purchase Stages*.

Liu, J., Dolan, P., and Pedersen, E. R. (2010). “Personalized news recommendation based on click behavior.” In *Proceedings of the 15th international conference on Intelligent user interfaces*, 31–40, ACM.

McCombs, M. E., and Shaw, D. L. (1972). “The agenda-setting function of mass media.” *Public opinion quarterly*, *36*(2), 176–187.

Milosavljevi, M., and Vobi, I. (2019). “Human still in the loop.” *Digital Journalism*, *7*(8), 1098–1116.

Nechushtai, E., and Lewis, S. C. (2019). “What kind of news gatekeepers do we want machines to be? filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations.” *Computers in Human Behavior*, *90*, 298–307.

Nicas, J. (2018). “Apple newss radical approach: Humans over machines.” *New York Times*, https://www.nytimes.com/2018/10/25/technology/apple--news--humans--algorithms.html.

Oestreicher-Singer, G., and Sundararajan, A. (2012). “The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets.” *Management Science*, *58*(11), 1963–1981.

Schaefer, M., and Sapi, G. (2020). “Learning from Data and Network Effects: The Example of Internet Search.” *DIW Discussion Paper 1894*.

Schepp, N.-P., and Wambach, A. (2016). “On big data and its relevance for market power assessment.” *Journal of European Competition Law & Practice*, *7*(2), 120–124.

Sen, A., and Yildirim, P. (2015). “Clicks bias in editorial decisions: How does popularity shape online news coverage?” *Available at SSRN 2619440*.

Senecal, S., and Nantel, J. (2004). “The influence of online product recommendations on consumers online choices.” *Journal of retailing*, *80*(2), 159–169.

Sinha, R., and Swearingen, K. (2011). “Comparing recommendations made by online systems and friends.”

Sun, T., Yuan, Z., Li, C., Zhang, K., and Xu, J. (2020). “The value of personal data in internet commerce: A high-stake field experiment on data regulation policy.” *Available at SSRN 3566758*.

Sundararajan, A. (2008). “Local network effects and complex network structure.” *The BE Journal of Theoretical Economics*, *7*(1).

Susarla, A. (2019). “The new digital divide is between people who opt out of algorithms and people who don’t.” *TheConversation.com*, https://tinyurl.com/y2ochy7z.

Tucker, C. (2019). "Digital data, platforms and the usual [antitrust] suspects: Network effects, switching costs, essential facility." *Review of Industrial Organization*, *54*(4), 683–694.

Valavi, E., Hestness, J., Ardalani, N., and Iansiti, M. (2020). "Time and the value of data." *Working Paper.*

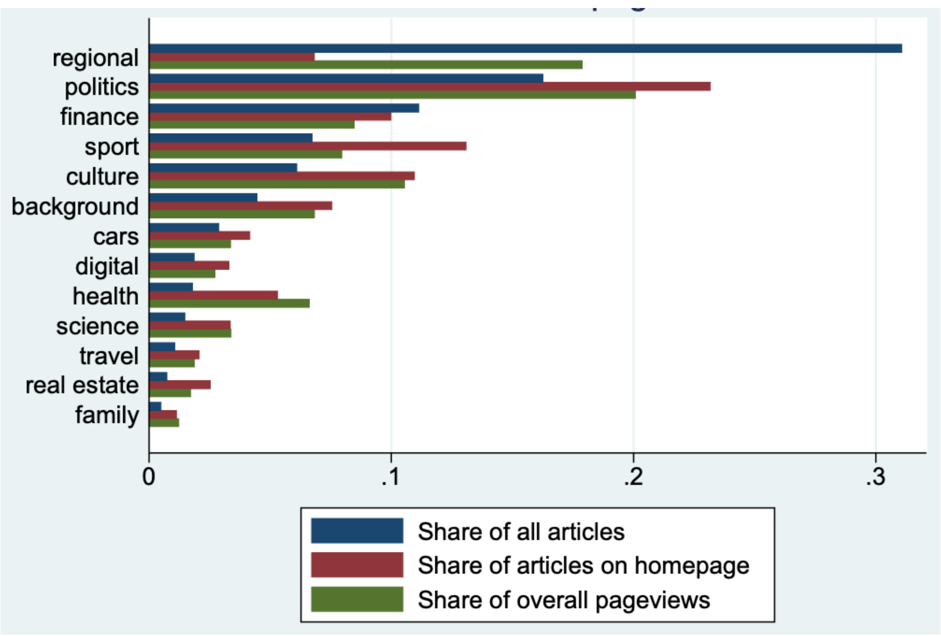
Wagner, G. G., Frick, J. R., and Schupp, J. (2007). "The German Socio-Economic Panel Study (SOEP)-Evolution, Scope and Enhancements." *Schmollers Jahrbuch*, *127*(1), 139–169.

Waldfogel, J. (2017). "How digitization has created a golden age of music, movies, books, and television." *Journal of Economic Perspectives*, *31*(3), 195–214.

Appendix

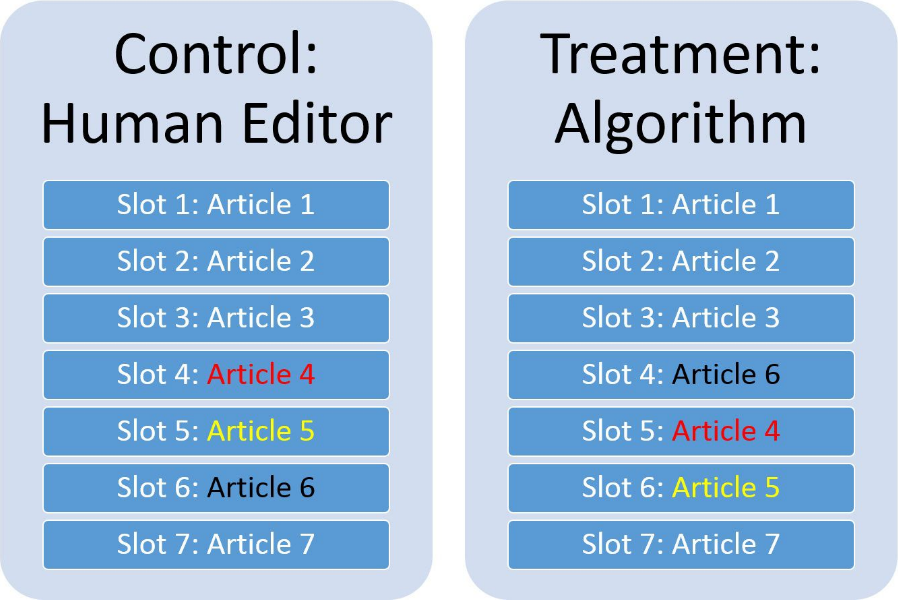
Figures and Tables

Figure 1: News Supply, Editorial Decisions and Clicks



The figure shows the supply of news articles by category in blue along with the share of categories that make it to the home page of the website in red. The green bars show the share of overall pageviews garnered by articles in that category.

Figure 2: Layout of Homepage with Control and Treatment



The figure 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. Example shown here: Algorithm selects item on slot 6 to moved upwards.

Table 1: Randomization Check

	(1) Control	(2) Treatment	(3) Overall	(4) (1) vs. (2) p-value
% Days active	9.2478	9.2524	9.2501	0.7204
	(0.0090)	(0.0090)	(0.0064)	
Total clicks	47.9718	48.0579	48.0150	0.5825
	(0.1107)	(0.1108)	(0.0783)	
Clicks/Day	3.6785	3.6861	3.6823	0.2696
	(0.0048)	(0.0048)	(0.0034)	
Clicks/Work Hours	24.5013	24.5463	24.5239	0.5774
	(0.0571)	(0.0572)	(0.0404)	
German Clicks	42.4909	42.5514	42.5212	0.6831
	(0.1048)	(0.1049)	(0.0741)	
N	1003285	1002509	2005794	

Column (4) provides the p-value in the difference in means between the treatment (column (2)) and control group (column (1)). The number of observations refers to individuals who we observe in the month before the experiment began. Percent days active refers to the percentage of days an individual was active in the month before the experiment. Total clicks refers to the number of clicks before the experimental period, clicks during the day, clicks during work hours and clicks from individuals browsing from within Germany are also based on numbers from the pre-experimental month.

Table 2: Summary Statistics

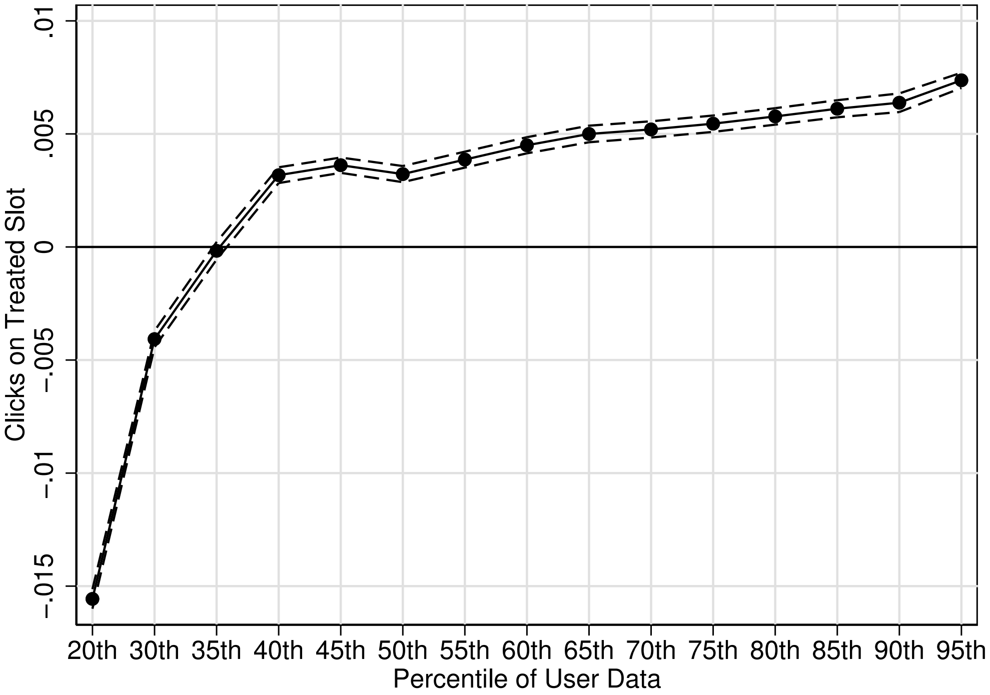
VARIABLES	Obs.	Mean	Std. Dev.	Min.	Max.
Hits on Slot 4	154,616,084	0.0277	0.185	0	110
Hits on Other Slots	154,616,084	0.7262	1.335	0	2809
Cosine Distance (Pre-Treatment Users)	63,713,363	0.314	0.230	0.274	1
User-HHI (Pre-Treatment Users)	63,713,363	0.182	0.180	0.0149	1
Hits on Slot 4 (New Year Bug)	10,366,108	0.0242	0.165	0	54
Number of Visits (Logarithm)	154,616,084	2.805	2.178	0	11.72

Table 3: Baseline and Scale Effects

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Slot=4	Slot=4	Slot=4	Slot≠4	Slot=4	Slot=4
Treatment	-0.0003***	0.001***	-0.008***	-0.066***	0.0086***	0.0057***
	(0.00009)	(0.00004)	(0.00008)	(0.00052)	(.00010)	(0.00031)
Treatment × Prior Visits			0.003***	0.022***		0.00047***
			(0.00003)	(0.00017)		(0.000067)
Treatment × Cosine Distance					-.0105***	-0.0148***
					(0.0002)	(0.0006)
Treatment × Prior Visits × Cosine Distance						0.0017***
						(0.0001)
Day FE	No	Yes	Yes	Yes	Yes	Yes
Individual FE	No	Yes	Yes	Yes	Yes	Yes
Observations	154616084	137689847	137689847	137689847	63341793	63341793
R-Squared	0.000	0.105	0.106	0.230	0.0258	0.0257

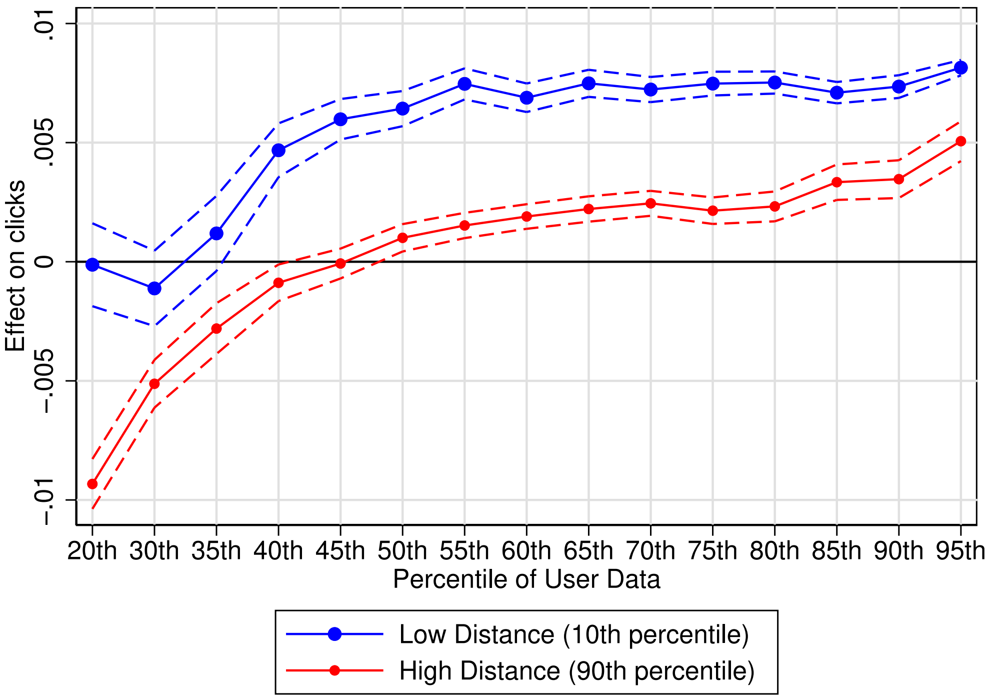
The dependent variable is the number of clicks on Slot 4 in columns (1), (2), (3) and (5), and clicks on other slots in column (4). *Prior Visits* is the user's log number of prior visits to the homepage since the beginning of the experiment. Cosine Distance is computed on the pre-experimental sample of users. Some of the baseline interactions are suppressed for better legibility. The unit of observation is user-session. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Figure 3: Diminishing Returns to Data



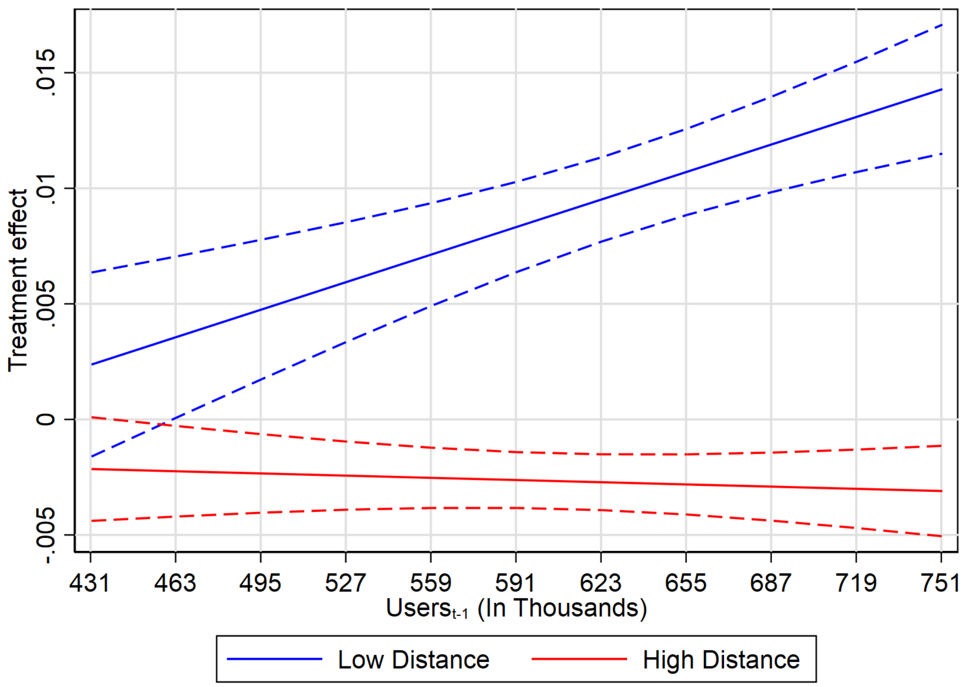
The figure plots the coefficients of the treatment effects associated with algorithmic recommendations relative to the human editor along with 99% confidence intervals based on the different data bins specified in regression (2). The vertical axis captures the magnitudes of the treatment effects of algorithmic recommendations relative to the human with the horizontal axis capturing the percentiles of the number of visits of an individual user. The values before the 20th percentile are suppressed because of limited variation in prior visits of a user. The median number of visits is 16. The dependent variable is the number of clicks on slot 4. The unit of observation is user-session. The sample used includes all individuals observed during the experimental period. The unit of observation is user-treatment.

Figure 4: Returns to Data and Across-User Consumption Diversity



The figure plots the coefficients of the treatment effect associated with different levels of user-heterogeneity (cosine distance) over the experimental period along with 99% confidence intervals. The vertical axis captures the magnitudes of the coefficients with the horizontal axis capturing the percentiles of the number of visits of an individual user. The values before the 20th percentile are suppressed because of limited variation in prior visits of a user. The median number of visits is 16. The dependent variable is the number of clicks on the treatment slot. The unit of observation is user-session and sample includes only those individuals who we observe in the pre-experimental period as well.

Figure 5: Data Network Effects and Algorithmic Recommendation



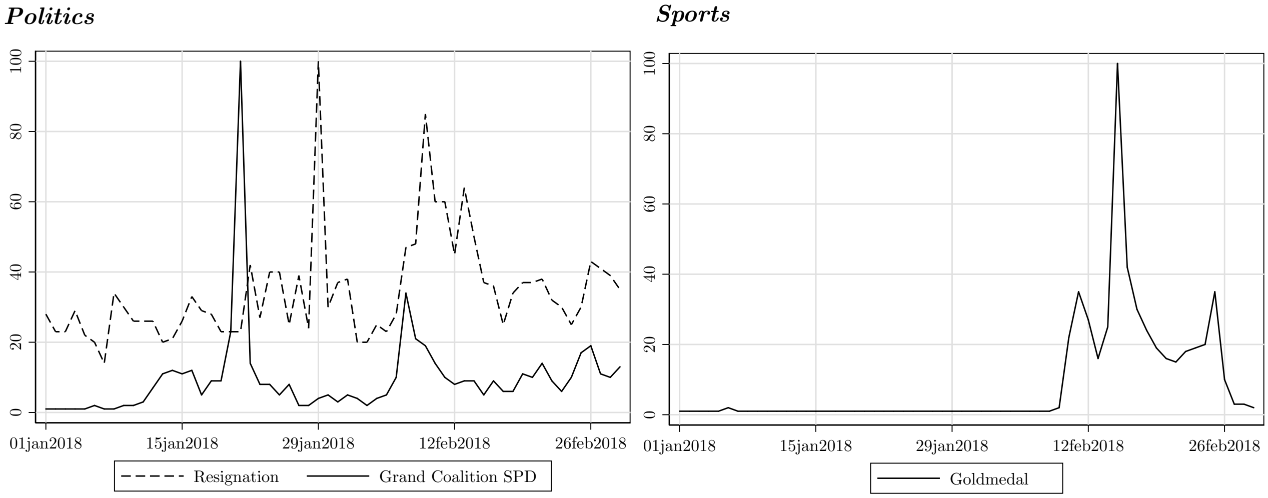
The figure plots the coefficients of the treatment effect associated with different user base size at different levels of data heterogeneity based on users' browsing behavior, over the experimental period along with 90% confidence intervals. The vertical axis captures the magnitudes of the coefficients with the horizontal axis capturing the number of users on the website on the previous day. The dependent variable is the number of clicks on the treatment slot. The unit of observation is user-session.

Table 4: Stock and Flow of Data: A Natural Experiment

	(Dec-Jan)	(Dec-Jan)	(Bug Period)
	(1)	(2)	(3)
VARIABLES	Slot=4	Slot=4	Slot=4
Treatment	0.002***	0.002***	-0.009***
	(0.00006)	(0.00006)	(0.0002)
Treatment × New Year Bug	-0.005***		
	(0.00014)		
Treatment × New Year Bug Day Trend		-0.001***	
		(0.00003)	
Treatment × Visits Prior to New Year Bug			0.003***
			(0.00007)
Day FE	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes
Observations	71901149	71901149	9191465
R-squared	0.107	0.107	0.145

The dependent variable is the number of clicks on slot 4. The unit of observation is user-session. Columns (1) and (2) confine the sample to December 2017 and January 2018 to analyze the New Year coding bug. Column (3) contains only those users who visited the website during the bug period with the time-frame includes only the bug week. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Figure 6: Surprising News Events: Google Trends



Relative search volume for the terms “Rücktritt” (resignation), “Groko SPD” (Grand Coalition SPD) and “Goldmedaille” (Goldmedal) on Google in Germany, as reported by Google Trends. Spikes conincide with major news events.

Table 5: Surprising News and Algorithmic Performance

VARIABLES	(1) Slot=4 (Politics)	(2) Slot=4 (Politics)	(3) Slot=4 (Sports)	(4) Slot=4 (Overall)	(5) Slot=4 (Overall)
Treatment	0.00877*** (0.00005)	0.00776*** (0.00004)	0.00042*** (0.00007)	0.00112*** (0.00004)	0.00522*** (0.00012)
Treatment × GrandColSPD	-0.00106*** (0.00020)				
Treatment × Resignation		-0.00484*** (0.00016)			
Treatment × Goldmedal			-0.00837*** (0.00020)		
Treatment × HighVarianceDays				-0.00207*** (0.00009)	
Treatment × DemandVariance					-0.00142*** (0.00004)
Observations	37471493	61981416	24370860	137689847	137689847
R2	0.11716	0.11372	0.12710	0.10498	0.10499

The dependent variable is the number of clicks on politics articles on slot 4 in columns (1)-(2), the number of clicks on sports articles on slot 4 in (3), and the overall number of clicks on slot 4 in columns (4)-(5). The unit of observation is user-session. The number of observations includes all individuals observed during the experimental period in columns (4) and (5), in column (1) during January 2018, and in column (2) during January and February 2018 and in column (3) during February 2018. News events are defined as indicating the days of the respective spikes in Figure 6. Columns (4)-(5) use general measures of demand variance. In column (4) we focus on a dummy indicating days with high demand variance. In column (5) we use the continous measure of demand variance: 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. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 6: Robustness: Alternative Functional Forms

VARIABLES	Log(Clicks)
	(1)
	Slot=4
		Prob(Clicks)
		(2)
		Slot=4
			Poisson-Clicks
			(3)
			Slot=4
				Neg. Bin.-Clicks
				(4)
				Slot=4

Treatment	0.0004***	0.0004***	0.0248***	0.0147***
	(0.00003)	(0.00004)	(0.00143)	(0.00116)
Individual FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Observations	137689847	137689847	84421660	84421660
R-squared	0.104	0.103	-	-

The dependent variable is log(1+number of clicks on Slot 4) in column (1), it is the probability of any click on Slot 4 in column (2) and we use the number of clicks on Slot 4 in (3) where we use a Poisson model while in column (4) we use a Negative Binomial model. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period for columns (1)-(4). The number of observations is smaller in columns (3) and (4) because observations without variation (all zeros) are explicitly dropped. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 7: Alternative Baseline Specification with Finer Fixed Effects

VARIABLES	(1) Slot=4	(2) Slot=4	(3) Slot=4	(4) Slot=4
Treatment	0.001***	0.002***	0.002***	0.002***
	(0.00004)	(0.00004)	(0.00004)	(0.00005)
Day FE	Yes	Yes	Yes	Yes
Fixed FE	User-Week	User-Day	User-Hour	User-Hour of Day
Observations	144835865	145273069	143403311	141598909
R-squared	0.117	0.152	0.131	0.171

The dependent variable is the number of clicks on Slot 4. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 8: Robustness: Alternative Variation and Rank Effects

	First Session of Hour	First Session of Morning	First Session of Afternoon	Treatment Group	Treatment Group
	(1)	(2)	(3)	(4)	(5)
VARIABLES	Slot 4	Slot 4	Slot 4	Slot 4	Slot 4
Treatment	-0.004***	-0.006***	-0.005***		
	(0.00032)	(0.00041)	(0.00040)		
Treatment × Prior Visits	0.001***	0.002***	0.002***		
	(0.00007)	(0.0001)	(0.0001)		
Rank				-0.002***	-0.005***
				(0.0002)	(0.0003)
Rank × Prior Visits				0.0002**	0.0012***
				(0.00008)	(0.00008)
Individual FE	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes
Observations	38982413	24324458	24943519	51,908,122	13,514,178
R-squared	0.037	0.051	0.049	0.102	0.147

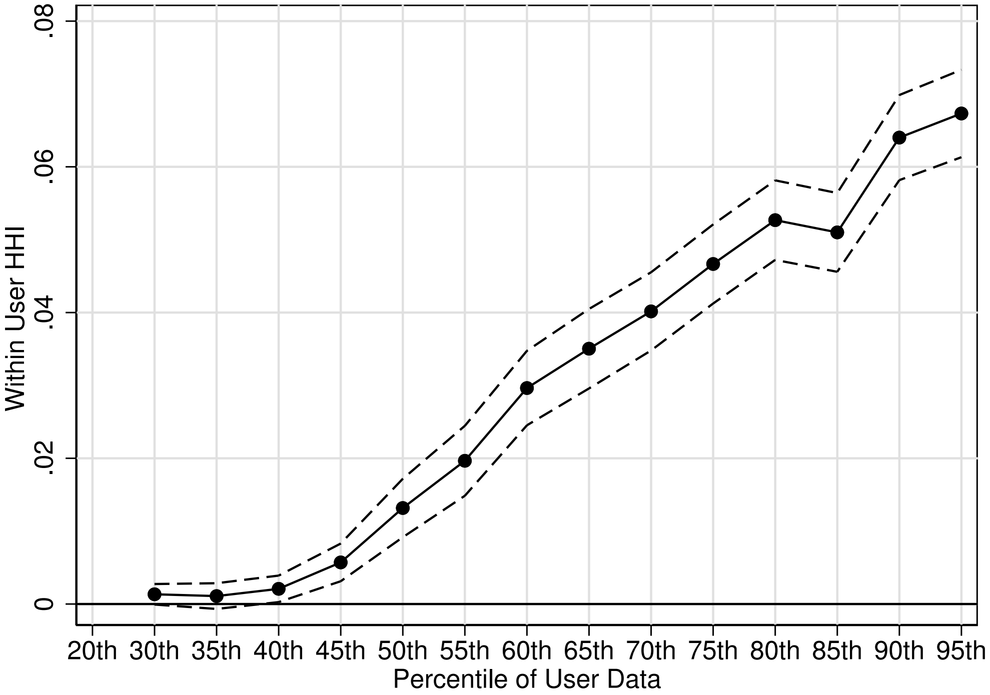
The dependent variable is the number of clicks on Slot 4. Column (1) uses data from the first session of every hour for a user, and column (2) uses data from the first session of the morning (6am to 12pm), and column (3) uses data from the first session of the afternoon (12 pm to 6 pm). Columns (4) and (5) use the sample of users only when they are in the algorithmic treatment group. Column (4) uses articles from all slots below slot 4 while column (5) uses articles only from slots below slot 14. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 9: Algorithmic Recommendation and Within-User Consumption Diversity

	(1) Slot=4	(2) Slot=4	(3) Slot=4
Treatment	0.052***	0.046***	-0.046***
	(0.00061)	(0.00122)	(0.00142)
Treatment × Pre_User_HHI		0.065***	
		(0.00481)	
Treatment × Total Visits			0.022***
			(0.00034)
Individual FE	Yes	Yes	Yes
Observations	654594	387752	654594
R2	0.687	0.651	0.689

The dependent variable is the within-user HHI where a higher value measures a decrease in consumption diversity. The unit of observation is user-treatment. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Figure 7: Within User Diversity and Algorithmic Recommendation



The figure plots the coefficients of the treatment interacted with the number of visits (data bins) over the experimental period along with 99% confidence intervals. The vertical axis captures the magnitudes of the coefficients with the horizontal axis capturing the number of visits of an individual user. The dependent variable is the individual level HHI over the entire experimental period. The unit of observation is user-treatment. The value at the 20th percentile is suppressed because there is only one observation and the diversity measure would be equal to one by definition.

Table 10: Information Externalities: Political News and Reader Characteristics

VARIABLES	(1) Share Politics (Slot4)	(2) Share Politics (Slot4)	(3) Share Politics (Slot4)
Treatment	0.005*** (0.00003)	0.004*** (0.00009)	0.022*** (0.00094)
Treatment × No Desktop/Laptop	0.002*** (0.00006)		
Treatment × Extreme Vote		0.006*** (0.00043)	
Treatment × Voter Turnout			-0.022*** (0.00122)
Day FE	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes
Observations	137689847	132215191	132215191
R-squared	0.102	0.098	0.098

The dependent variable is the share of clicks on political stories displayed on Slot 4. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period. Sample only includes users within Germany in columns (2) and (3). Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

A Supplementary Appendix

Table A.1: Survey Items – Digital Literacy

(1) I use a computer at work. (*agree*/don't agree)

(2) I know how to code or have taken a computer science class. (*agree*/don't agree)

(3) What is HTTP? (a) Operating system, (b) physical parts of a computer, (c) *fundamental technology for communication in the WWW*, (d) I don't know.

(4) Which technology makes your transactions with online merchants secure? (a) Microsoft Windows Firewall (MWF), (b) Cookies, (c) *Secure Sockets Layer (SSL)*, (d) I don't know.

(5) What is "machine learning"? (a) software-technology for schools and universities, (b) software-technology based on rules, (c) *software-technology based on statistics*, (d) I don't know.

Cumulating the answers in *italics*, our index has a maximum score of 5. Our digital literacy score has a mean of 2.998, standard deviation 1.188, min 0 and max 5.

Table A.2: Survey Results – Correlation with Digital Literacy

VARIABLES	Digital Literacy	
No Laptop/Desktop	-0.422***	(0.107)
Age	-0.004	(0.004)
Income	0.076***	(0.026)
Education	0.455***	(0.051)
Observations	497	
R-squared	0.199	

The dependent variable is the digital literacy score as defined in Table A.1. White-robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.3: Online Experiment: Algorithmic and Editorial Recommendations

	(1)	(2)	(3)
VARIABLES	Treated Slot	Untreated Slot	All Slots
Personalized	-0.002	-0.000	-0.015
	(0.021)	(0.0069)	(0.016)
Curated	-0.001	-0.003	-0.004
	(0.020)	(0.007)	(0.015)
Untreated Slot × Personalized			0.017
			(0.0179)
Untreated Slot × Curated			0.002
			(0.0017)
Individual FE	Yes	Yes	Yes
Observations	4239	21195	25434
R-Squared	0.542	0.142	0.146

The dependent variable is the number of clicks on the treated slot in column (1), the untreated slot in column (2) and all slots in column (3). In the curated condition, we label the news story on slot 4 as "Recommended by the editor". In the personalized condition, we label the news story on slot 4 as "Recommended for you". The unit of observation is at the user-level. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Some Technical Details of the Algorithm

In this section, we provide a brief technical overview of the algorithm put forward in Liu et al. (2010) which the news outlet's data science team used as a baseline framework. The details sketched out below were used as a guiding framework by the team. Our communications with the team indicated that they improved upon some dimensions of the model to make it a better fit for their particular data and context. The recommendation algorithm uses a combination of personal and social data to predict a user's interest in a certain news category: her own click behavior, augmented by the current news trend proxied by the reading behavior of other users.

The interest of user i at time t in news articles of category c_j is captured by the probability she clicked on an article in that category: $interest^t_{i,j}(cat = c_j) = p^t(click_i|cat = c_j)$ which is computed using Bayes' rule:

$$\frac{p^t_{i,j}(cat = c_j|click_i) \times p^t(click_i)}{p^t(cat = c_j)},$$

where $p^t(cat = c_j|click_i)$ is the probability of the user's click being in category c_i which can be estimated from the distribution of clicks of the user across topics over time, which can be written as:

$$D(i,t) = \left(\frac{N^t_{i,1}}{N^t_{i,total}}, \frac{N^t_{i,2}}{N^t_{i,total}}, ..., \frac{N^t_{i,n}}{N^t_{i,total}} \right)$$

where $N^t_{i,total} = \sum_j N^t_{i,j}$ is the total number of clicks by user i in time period t. The probability that an article is in category c_j, $p^t(cat = c_j)$, is a result of the supply of news articles, but can be approximated by the overall demand for articles in that category, i.e. the click distribution of the population across topics $D(t)$. Finally, $p^t(click_i)$ is the probability that user i clicks on any article irrespective of the category and can be approximated by user i's total clicks relative to the population's total clicks in period t.

As new data arrives over time, the model's prediction of user i's interest in category j is updated by combining data from all available time periods, such that

$$interest^t_{i,j}(cat = c_j) = \frac{\sum_{\tau=1}^{t-1} \left(N^\tau_i \times interest^\tau_{i,j}(cat = c_j) \right)}{\sum_{\tau=1}^{t-1} N^\tau_i}$$

where N^t_i is the total number of clicks at time t.

Because news articles arrive frequently, and not all users have a long histories, the model would suffer from a set of 'cold start' problems when trying to predict a content categories to users. Hence, it additionally includes data on the aggregate reading behavior of other users.

As a remedy to the problem of new items arriving, we can augment the above described information filtering algorithm with social data. General interest in a news article of category c_j at time t can be defined as $interest^t_j(cat = c_j) = p^t(cat = c_j)$. With a large enough number of users, this can be approximated by the overall click distribution over a relatively short time period $t = 1, ..., \gamma$ (a few hours or a day). Under the assumption that general interest in category j as aggregated over the period γ is proportional to user i's interest in that category, it can be used to inform predictions about user i's probability to click on an article in category j in the near future. The augmented model can be written as

$$\begin{aligned} p^t_{i,j}(cat = c_j|click_i) &= interest^t_{i,j}(cat = c_j) \times G(interest^\gamma_j(cat = c_j)) \\ &= \frac{p^\gamma_j(cat = c_j) \times \Big(\sum_{\tau=1}^{t-1} \Big(N^\tau_i \times interest^\tau_{i,j}(cat = c_j) + G\Big)}{\sum_{\tau=1}^{t-1} N^\tau_i + G} \end{aligned} \tag{4}$$

where G is a weighting parameter that can be interpreted as simulating user i's clicks perfectly following the distribution of the current news trend. From equation 4 it becomes clear that the model's predictions will be entirely based on the current news trend when there is zero individual specific data available, and the model's predictions will increasingly be based on a combination of other peoples' reading behavior and user i's personal data $\sum N_i$ grows larger.

The model's predictions for each user and category are sorted to select the category with the highest predicted likelihood of clicking. The algorithm then selects an article in that category from the pool of articles that the human editor has selected to appear on the homepage at any given moment. Each user's reading behavior is continuously fed into the recommendation system and the prediction scores for each user and category are updated on a daily basis.