The Editor vs. the Algorithm: Targeting, Data and Externalities in Online News[*]

Jörg Claussen
LMU Munich[†]

Christian Peukert
CLSBE and ETH Zurich[‡]

Ananya Sen
MIT Sloan[§]

June 5, 2019

Abstract

We run a field experiment with a major news outlet to quantify the economic returns to data and informational externalities associated with algorithmic recommendation. Automated recommendation can outperform a human editor in terms of user engagement, though this crucially depends on the amount of training data. Limited individual data or breaking news leads the editor to outperform the algorithm. Additional data helps algorithmic performance but decreasing economic returns set in rapidly. Investigating informational externalities highlights that personalized recommendation reduces consumption diversity. 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 Marco Savioli, Doh Shin Jeon, Jacques Cremer and Lorien Sabatino as well as participants at the KOF-ETH Zurich Seminar, MIT CODE conference 2018, 16th Media Economics Workshop in Lecce, MIT IDE Annual Conference, Munich Beliefs about Society and Politics Workshop and the Workshop on the Economics of Data and AI at the European Commission for helpful suggestions. Peukert acknowledges support from FCT – Portuguese Foundation of Science and Technology for the project UID/GES/00407/2013 and FCT-PTDC/EGE-OGE/27968/2017.

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

[‡] Catolica-Lisbon School of Business and Economics; ETH Zurich, Center for Law and Economics, christian.peukert@gmail.com.

[§] MIT Sloan School of Management, ananyas@mit.edu.

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) technologies are starting to be utilized in a large number of industries. The increased market power of online platforms has often been attributed to the use of these techniques combined with access to a plethora of individual level data. This rise in concentration has caught the attention of policy makers who increasingly believe that 'scale effects' of individual level data can be a significant source of anti-competitive behavior. The regulatory environment in a number of regions (e.g. GDPR, California Consumer Privacy Act) is attempting to put restrictions on what firms can do with user specific information, which has brought such issues to the forefront of the discourse both from a competition policy as well as a consumer privacy perspective. These discussions are also related, more generally, to the issue of automation of tasks since how the gulf between human and algorithms might widen because of an algorithm's access to rich consumer information is at the core of the debate.[1] Moreover, it is unclear how humans would perform relative to algorithms in 'creative' industries (e.g. news and media industry) since the focus of automation has been on 'routine' tasks which involve minimal interpretation and subjective judgment.[2] In the case of online news, algorithmic recommendations might, additionally, lead to unintended consequences and a (socially) 'less desired' outcome if readers don't account for informational externalities of their own reading behavior. This assumes greater significance if readers confine themselves into echo chambers with algorithms trained on prior individual level data reinforcing this phenomenon (Gentzkow, 2018).

To explore these interrelated issues of algorithms, economic returns to data and informational externalities, we partner with a major German news outlet to carry out a field experiment. The home page of the news outlet's website is always curated by a human editor. At each point in time, N articles are featured on the homepage. In general, any user that arrives at the homepage sees the same content in the same place. In the experiment, every time a user visits the homepage, she is randomly assigned to a control or treatment condition. If a user is assigned

[1] A large number of online platforms such as Apple, Spotify, Google and Amazon are under scrutiny due to competition policy concerns partially because of their ability to leverage consumer level data to better 'match' their subjective preferences.

[2] Within the news industry, in particular, editorial decisions necessarily involve subjective judgments about 'newsworthiness' of stories. This inherent subjectivity over the choice of news stories could explain the debate in the industry between using human 'curators' instead of opting for an automated system. In fact, Apple News recently hired a sleuth of human curators instead of having algorithms choose the news for its customers. For more on this see `https://www.nytimes.com/2018/10/25/technology/apple-news-humans-algorithms.html` as well as `https://www.forbes.com/sites/stevenrosenbaum/2015/07/26/the-curation-explosion/#4befb785409c` for a broader discussion about curation in the industry.

to a control condition then all the articles she observes on the homepage are the ones curated by the human editor. In the treatment condition, we customize the homepage by implementing a (widely used) recommendation algorithm, trained on fine grained individual level data, decide which of the N articles to be placed on a specific (fixed) slot $n = 4$.[3] In this setting, we first ask whether algorithmic recommendations can outperform a human editor in terms of user engagement (e.g. clicks) and under which conditions the human editor can win against the algorithm. This can be especially pertinent in the context of online news since editorial experience in identifying the 'importance' of news stories is said to be crucial for a successful outlet. More generally, we investigate the economic returns to data and how the effectiveness of the algorithmic recommendation improves due to more training data, relative to the human editor. We analyze how algorithmic recommendation performs as the same user visits the website repeatedly providing variation in the amount of user-specific information that is used to train the recommendation system. Finally, we analyze the potential information externalities of such algorithmic recommendations. We construct measures of consumption diversity across different news categories, test whether it is impacted by personalized algorithmic recommendations and analyze which reader characteristics might be driving this behavior.

We find two broad set of results. First, the baseline model-free evidence shows that on average, the human editor outperforms the algorithm in terms for reader clicks. Introducing individual fixed effects, which eliminates the impact of one time visitors, flips this result with the algorithmic recommendation doing better relative to the editor. Additionally, the human editor performs better than the algorithm on days with fast developments in breaking news events. This suggests that the human editor might be better at predicting the taste of the average reader in the population which in turn implies that a combination of the human and algorithmic editor might provide the biggest payoff to the firm. More generally, we find that after about 6-10 visits by an individual, the algorithm consistently outperforms the human editor as each visit allows the algorithm to get more detailed data. While data helps algorithmic performance, we show that there are decreasing *economic returns* to data which set in rapidly with an additional user visit leading to smaller improvements in algorithmic performance. These results imply that there might be limited strategic advantages for a firm by simply having access to individual level data. Additionally, if privacy concerns lead to limits on retention then this

[3] The algorithm will use the pool of articles which are listed below slot 4 to "push up" based on prior reading behavior. See the section below for more details on the algorithm used.

should not have too big an impact on algorithmic performance.

Second, we find that algorithmic recommendation reduces the consumption diversity by users when they are in the treatment group relative to the control group. This reduction in consumption diversity spills over onto other slots as well. This implies that users tend to click more on other slots related to similar topics while in the treatment condition. Using pre-experimental data we show that, for example, readers who had a higher share of politics consumption increase it even further during the experiment. Additionally, we show that proxies of digital literacy[4] and extreme political views are associated with a tendency to reduce consumption diversity in line with popular discourse.[5] These results speak directly to the recent conversation about the impact of filter bubbles and echo chambers because of algorithmic recommendations.

Our findings contribute to several streams of the literature. First, we complement a few existing studies which look at the scale effects of data on measures of algorithmic performance. Chiou and Tucker (2017) analyze a policy change in Europe which reduced the time window search engines could retain individual user data and find that it did not affect the accuracy of search results related to the news stories of the day. Schaefer et al. (2018), 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. Similarly, Bajari et al. (2018) analyzing product forecast accuracy using data from Amazon find improvements in forecast accuracy with certain types of additional data. They also note how there are very few existing studies which truly test the 'scale effects of data' hypothesis. They acknowledge the limitations of their own 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 scale effects of data with variation coming from a randomized experiment. Our results on economic returns to data provide a nuanced view which might reconcile the effects found in this literature.

It is crucial to note that such studies are important because of their focus on the *economic returns* to data rather than one which confines itself solely to improvement in algorithmic precision due to additional data. The response of prediction accuracy of ML models to additional data is governed to a great extent by the underlying statistical model. In this paper, by focus-

[4] We carry out a survey on a representative German sample to determine correlates of digital literacy which we can map back into our browsing data. The most significant correlate is the lack of ownership of a laptop or a desktop - a variable which we also see in our main dataset. See Table A.3 and A.4 in the Supplementary Appendix for details.

[5] See for example Susarla (2019).

ing on economic returns, we are also attempting to map data into reader preferences and firm revenue. A priori, one could expect discontinuities, threshold effects and increasing returns as we map additional data for the algorithm into economic outcomes. This has been a first order concern for policy makers.

We also contribute to the literature investigating which tasks might be suitable for automation and where humans would still hold an edge in the foreseeable future. Agrawal et al. (2018) highlight that the main area machine learning and AI will reshape tasks are those which involve prediction while humans will hold the key in those which require subjective judgment.[6] 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. We add to this mixed picture by showing that a combination of human and algorithms might best serve the strategic interests of the firm, especially when subjective judgments are to be made as is the case in determining 'newsworthiness' of stories.[7]

Analyzing the externalities of personalized news recommendations also contributes to the literature on the role of the internet and the resulting echo chambers in increased political polarization (e.g. Gentzkow and Shapiro, 2011; Boxell et al., 2017; Bakshy et al., 2015). The fact that personalized algorithmic recommendation can lead to a reduction in diversity of news consumption away from political information goes to the core of the issue of divergence between individual and social preferences. To the best of our knowledge, ours is the first paper to analyze diversity in individual news consumption which goes beyond descriptive analysis and speaks to the issues raised by Gentzkow (2018).[8]

2 Background and Experimental Setting

Our partner news outlet is one of the largest players in the German news market with over twenty million monthly unique visitors to its website. It is similar to a publication like the Wall

[6] Relatedly, Brynjolfsson and Mitchell (2017) emphasize that no job will be completely automated though some tasks associated with different job will be "suitable for ML".

[7] There are other studies (e.g. Shichor and Netzer, 2018) which train machine learning models to mimic human decisions but do not have a randomized experiment to enable clean causal analysis. See Mullainathan and Spiess (2017) for more details.

[8] More generally, we contribute to the literature about the intended and unintended effects of recommendation tools in news aggregation (George and Hogendorn, 2013; Calzada and Gil, 2016; Oh et al., 2016; Athey et al., 2017; Chiou and Tucker, 2017), e-commerce settings (e.g. Oestreicher-Singer and Sundararajan, 2012b,a; Hosanagar et al., 2014), and online advertising (e.g. Lambrecht and Tucker, forthcoming).

Street Journal in size and influence and like other major news outlets, our partner gets a large share of its revenue from advertising which makes reader engagement (e.g. clicks) crucial for its financial health. More generally, 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, has recently started experimenting with personalization of an individual reader's newsfeed only based on geographical location.[9]

Our randomization ensures that if a user is assigned to the control group, then she sees the homepage curated by the human editor which involves no personalization and anyone assigned to the control group at a particular instant sees the same layout. 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 sees the same "ordinal ranking" as the control group except for this change. The Data Science team implemented an improved version of a widely used machine learning model (based on hybrid methods), including Google News, which is trained on fine grained data about the past reading behavior of each individual user as well as news reading trends of other users.[10] The features used to train the model include detailed article-level information including keywords and tags. A user is identified based on a unique cookie ID. Given prior reading behavior, the model's output is a prediction score of how likely the user would be to click on an article in a given category. The algorithm then selects an article in the category with the highest likelihood from the pool of articles that the human editor has selected to appear on the homepage at any given moment. If the highest click probability article is already on a slot above (n=1, 2 or 3), then the system chooses the next best. In essence, the algorithm works by rearranging the human editor's ranking of articles on slot 4 and correspondingly moving other articles up or down in the ranking. Each user's reading behavior is continuously fed into the recommendation system and the prediction scores for each user and category are updated. If a user has no prior reading behavior, then the system assigns a recommendation that is based

[9] See `https://www.nytimes.com/2017/03/18/public-editor/a-community-of-one-the-times-gets-tailored.html` for more on the experiments underway and the strategy for the future.

[10] The algorithm implemented by the firm's data science team had the method put forward in Liu et al. (2010), developed by Google engineers for Google News, at the core of it. The method uses a combination of past reading behavior of the individual user as well as a collaborative filtering mechanism to provide recommendations. This core model was then updated with some additions to ensure the best performance.

on the collaborative filter driven by features across other users' current reading behavior – the algorithm is not (necessarily) replicating the human editor's choice for slot 4. The randomization is at the user-session level such that when the user is inactive for thirty minutes and/or reloads the homepage, the randomization takes places again. This allows us to utilize user fixed effects and cleanly identify time varying coefficients, such as the effect of the amount of user-specific data on algorithmic effectiveness. The experiment was carried out from December 2017 to May 2018.

3 Empirical Framework

Our baseline specification links reader engagement on the website to whether she was in the treatment or control group:

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

The unit of observation in our empirical analysis is user-session. We define a session to include all clicks that a user makes until there is inactivity for thirty minutes. We focus on $Clicks_{is}$ as the main dependent variable of interest which represents the number of clicks by user i in session s.[11] We distinguish between clicks that originate from the treatment slot on the homepage (*Slot=4*) and other slots on the homepage (*Slot≠4*). Our main independent variable of interest is $Treatment_{is}$, which is whether user i was randomly assigned to the treatment group (algorithmic recommendation) in session s or if the user was in the control group (human curation). Theoretically, we should expect δ to be positive and statistically different from zero if the algorithm performs better than the human editor. We are also interested in clicks on other articles and in total clicks in a session, though the theoretical prediction for these are ambiguous. Even if the algorithmic recommendation outperforms the human editor on *Slot*= 4, it will depend on how attention spills over to other articles to determine whether there is a cannibalization or expansion effect overall. We include a time-trend γ_{τ} to control for events affecting all users, potentially through the news cycle. We use a time trend instead of day fixed effects to ensure faster estimation on such a large dataset. Our results are qualitatively and quantitatively the same when we include day fixed effects instead of a time trend.[12] User fixed effects μ_i capture

[11] As a robustness check, we also analyze our baseline results using the logarithm of clicks as the dependent variable.

[12] See Table A.1 (columns (4)-(6) to see the statistical and economic significance being almost identical to our baseline estimates.

time invariant differences in reader preferences over content. In our setting, introducing user fixed effects will eliminate the impact of one time visitors to the website, something we explore in detail below. We cluster standard errors at the individual level to account for serial correlation of user preferences over content.

4 Baseline Results and Scale effects of Data

4.1 Benchmark Results

Table 1: Randomization Check and Model Free Evidence

Panel A: Randomization Check

	(1) Control	(2) Treatment	(3) Difference((2)-(1))	(4) Std. Error	(5) Observations
Percent days active	0.3080	0.3082	0.0002	0.0004	2,004,597
Total clicks (norm.)	0.0393	0.0394	0.0001	0.0001	2,004,597
Clicks/Day (norm.)	0.0910	0.0911	0.00018	0.00012	2,004,597
Clicks/Work hours	0.5076	0.5079	0.0003	0.0005	2,004,597
Clicks from Germany	0.8832	0.8825	-0.0007	0.0004	2,004,597

Panel B: Model-Free Evidence

	(1) Control	(2) Treatment	(3) Difference((2)-(1))	(4) Std. Error	(5) Observations
Hits on Slot 4	0.0279	0.0276	-0.0003	0.0000	154,616,084
Hits on Other Slots	0.6902	0.6649	-0.0253	0.0002	154,616,084
Total Hits	0.7625	0.7438	-0.0187	0.0002	154,616,084

Standard errors in parentheses clustered at the user level. Column (3) measures the difference in means between the treatment and control group. The number of observations refers to individuals who we observe in the month before the experiment began in Panel A. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period in Panel B.

We first check the validity of our randomization procedure. In Table 1, we analyze the assignment of individuals into treatment and control groups 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. As can be seen, the sample is well balanced across all the observables, indicating that our randomization has worked in the desired manner.

Next, we analyze the impact of the treatment descriptively. Model free evidence in Panel B of Table 1 does not suggest that the algorithm outperforms the human editor in terms of user clicks. 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. Similarly, clicks on other slots also reduce in a significant manner. The fact that clicks on slot 4 and the neighboring slots move in the same direction with the treatment suggests that personalized recommendation may lead to a positive attention spillover to the other slots and does not cannibalize their clicks. While this result points to the inability of the algorithm to predict user preferences better than the human editor, we must exercise a bit of caution since this sample includes a sizeable number of visitors who arrive on the website only a few times, for whom the ML model has limited prior data.[13] To explore this issue further, we turn to regression analysis.

Results from an OLS estimation of equation 1 in Table 2 paints a more nuanced picture. In column (1), when we have user-fixed effects, we find that clicks that originate from slot 4 on the homepage increase by about 4% when it features a personalized recommendation, compared to the selection by the human editor.[14] This regression eliminates the impact of users who visit the website for only one session since their effect on τ is absorbed by the user fixed effects. In column (2), we look at some of the indirect effects that the experiment may have, to find that clicks to all other slots on the homepage increase by 1%. This suggests that the personalized recommendation has positive attention spillovers on the neighboring slots and does not cannibalize clicks that originate from the manually curated part of the homepage. The result is very similar in column (3), where we study the effect of getting an algorithmic recommendation on total clicks with a positive and significant effect of about 1% as well.

4.2 Scale Effects of Data and Algorithmic Performance

The fact that algorithmic recommendation might perform better with more data seems to be implied by our baseline results. Next, we go on to explore heterogeneity in the treatment effect, testing for scale effects of 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 2 with the number of past visits, i.e. the number of times user i

[13] These could be individuals that indeed only visit the outlet once, but also users that arrive without a cookie.

[14] Effect sizes are reported as relative to the baseline, i.e. the sample average.

Table 2: Baseline and Scale Effects

VARIABLES	(1) Slot=4	(2) Slot≠4	(3) Total	(4) Slot=4	(5) Total
Treatment	0.001	0.005	0.007	-0.00772	-0.0467
	(0.00004)	(0.00029)	(0.003)	(0.000)	(0.000)
Treatment × Prior Visits				0.00287	0.0190
				(0.000)	(0.000)
Constant	0.0266	0.582	0.768	0.0279	0.763
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Day Time Trend	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes
Observations	154,616,084	154,616,084	154,616,084	154,616,084	154,616,084
R-squared	0.147	0.276	0.280	0.181	0.141

Robust standard errors in parentheses clustered at the user level. The dependent variable is the number of clicks on Slot 4 in columns (1) and (4), total clicks in the session in (2) and (4) and clicks on other slots in column (3). The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period.

has visited the website since November 2017 up until that session. In results reported in columns (4) and (5) of Table 2, we find that clicks to articles on the treatment slots as well as overall clicks increase with the number of prior visits, i.e. as more information becomes available to the algorithm.

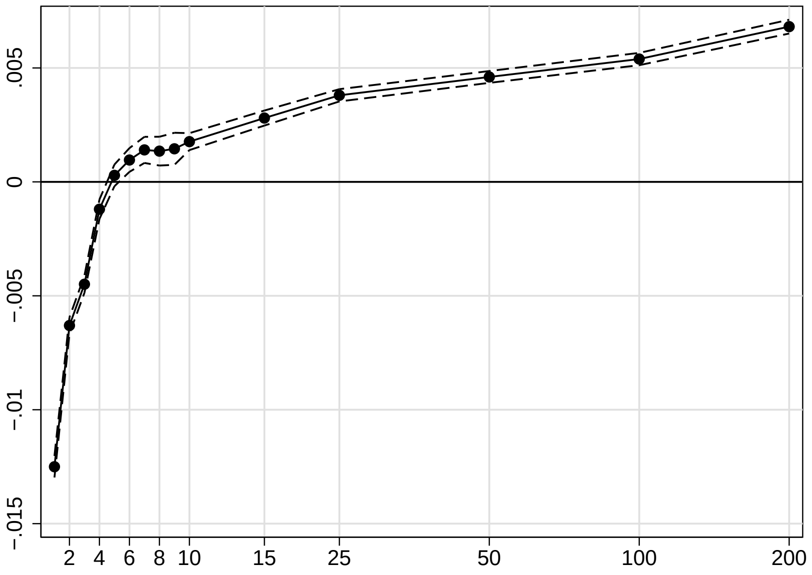
The above results, while illustrative, are still restrictive in analyzing the returns to data since we impose that engagement responds to prior data in a linear fashion. We adopt a more flexible approach by running the same regression but looking at finer data bins based on the number of past visits.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

$$Clicks_{is} = \alpha + \delta_1 Treatment_{is} + \sum_q \delta_q (Treatment_{is} \times PriorVisits_q) + \gamma_\tau + \varepsilon_{is} \qquad (2)$$

$$\forall q \in (1, 2, 3, ..., 9, 10-14, 15-24, 25-49, 50-99, 100-199, \geq 200).$$

The results in Figure 1 provide an insightful overview. Initially, when there is limited data for the algorithm then, as we noted above, the human editor outperforms the algorithm. This figure shows that when the algorithm has up to 5 visits per user then, the human has a comparative advantage. Around the threshold of ten visits, there is no (economically) significant difference between the human and algorithm performance. The gap between human and algorithmic performance gets wider, in favor of the algorithm, as more data is accumulated on past user behavior. Interestingly, we see that this gap levels off and stays the same beyond a threshold,

Figure 1: Decreasing Returns to data



The figure plots the coefficients δ_q along with confidence intervals based on the different data bins specified in regression (2). 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 number of clicks on slot 4. The unit of observation is user-session. The number of observations includes all individuals observed during the experimental period.

which is after a user has visited the website about 50 times previously. As can be seen from the figure, beyond that level of past usage, the impact on treated users clicking on the direct slot stays at similar levels of economic significance even though there might be some statistically significant differences.[15] Moreover, it is insightful to see that the returns to data results in a smooth curve without any obvious discontinuities, threshold effects or step functions.

If the human editor gains a competitive advantage over the algorithm because of limited data then, intuitively, we should also observe this phenomenon in the case of big breaking news event days. Due to limited data on big breaking news events, it can be envisaged that human editors are better at forecasting the 'newsworthiness' of a big developing story. We explore this dimension by analyzing 'surprising' developments related to the formation of the coalition between parties after the German federal elections over a period of December 2017 – February 2018. In column (1) of Table 3, based on events of 18th and 19th December 2017, we find that the editor beats the algorithm since the interaction term is negative and significant for clicks on slot 4. We repeat this exercise for the biggest 'surprising' events in January and February

[15] Algorithmic performance remains at similar economic levels even when we extend the series with finer intervals.

Table 3: Breaking News and Algorithmic Performance

VARIABLES	(1) Slot=4	(2) Slot=4	(3) Slot=4	(4) Total	(5) Total	(6) Total
Treatment	0.002 (0.00011)	0.004 (0.00009)	0.002 (0.00013)	0.006 (0.00078)	0.001 (0.00060)	0.008 (0.00073)
Treatment × News (Dec)	-0.004 (0.00024)			-0.009 (0.00192)		
Treatment × News (Jan)		-0.006 (0.00018)			0.001 (0.00143)	
Treatment × News (Feb)			-0.006 (0.00027)			-0.017 (0.00212)
Day Time Trend	Yes	Yes	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	27,889,311	42,258,526	27,486,627	27,889,311	42,258,526	27,486,627
R squared	0.175	0.174	0.214	0.286	0.321	0.365

Robust standard errors in parentheses clustered at the user level. The dependent variable is the number of clicks on slot 4 in columns (1)-(3) and total clicks in (4)-(6). The unit of observation is user-session. The number of observations includes all individuals observed during the experimental period in the particular month of breaking news considered.

related to the coalition talks to find very similar results in columns (2) and (3).[16] This effect also spills over to overall clicks in columns (4) and (6).

Overall, the algorithm outperforms the human editor when it has access to sufficient data, though in the early stages, the human is better at predicting the average taste of the readers. Hence, the optimal strategy for the news outlet seem to be to employ a combination of the algorithm and the human to maximize user engagement. This exercise sheds some light on the policy debate about data retention and firm performance. In particular, more individual level data can help firms gain a competitive advantage but we also see that decreasing economic returns set in quickly. The exact thresholds will, presumably, vary across different contexts and algorithms. This result, though, is broadly in line with the result of Chiou and Tucker (2017) where they show that a reduction in data retention doesn't affect search engine performance. Our results also suggest that legislation put forward by various institutions, including the European Commission on the amount of personal data retention by firms might not erode the competitive edge of firms in a significant manner since adverse consequences on consumer engagement and therefore firm performance would be limited.

[16] Our results are robust to alternative time thresholds for these events. For example, instead of looking at only the 18th and 19th of December, our results are unchanged if add the 17th of the month as well. This holds for other event months as well. See table A.2 for more details.

5 Information Externalities in Algorithmic Recommendations

The news is different from a standard product because of its public good nature. In particular, the algorithm is trained on prior individual level data, which is 'biased' towards personal preferences and could be at odds with "socially optimal" reading behavior.[17] 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 that go beyond the firm's intentions. We will analyze how algorithmic recommendations might have affected browsing behavior across different types of articles over the experimental period.

We use the Hirschman-Herfindahl Index (HHI) measure of consumption shares across different topics at the individual user level. 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. 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 4 show that the HHI increased when the users were in the treatment group relative to the control which means that the recommendation algorithm leads users to find similar topics to those recommended. This holds even when we focus only on articles read in the non-treatment slot (column 2). Overall, this implies that there was a reduction in the diversity of topics read on the treated slot which spilled onto to other slots as well. The magnitudes imply that there was a reduction in user level HHI by 5% for slot 4 and by 0.5% in terms of spillovers to other slots.[18] To dig deeper, we use pre-experimental browsing behavior for individuals who we also observe before the experiment to assess how their consumption diversity is affected by personalized recommendations. Focusing on political stories, columns (3) and (4) show that individuals who had a higher share of politics consumption in the pre-experiment period have an even higher share during the experiment due to the treatment.

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

[17] 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.

[18] 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. See `http://nymag.com/intelligencer/2014/02/why-comcasttime-warner-cable-should-be-blocked.html`.

Table 4: Algorithms and Information Externalities

Panel A: Automated Recommendations and Consumption Diversity

VARIABLES	(1) User HHI (Slot4)	(2) User HHI (Other)	(3) (Post) Politics (Slot 4)	(4) (Post) Politics (Other)
Treatment	0.049 (0.001)	0.005 (0.001)	0.002 (0.00057)	0.025 (0.00109)
Treatment x (Pre) Politics			0.002 (0.00042)	0.003 (0.00083)
Individual FE	Yes	Yes	Yes	Yes
Observations	2,446,445	23,426,694	63,706,396	63,706,396
R squared	0.12	0.64	0.03	0.15

Panel B: Consumption Diversity and Reader Characteristics

VARIABLES	(1) Share Politics (Slot4)	(2) Share Politics (Slot4)	(3) Share Politics (Slot4)
Treatment	0.006 (0.00003)	0.005 (0.00003)	0.022 (0.00097)
Treatment × No Desktop/Laptop	0.002 (0.00008)		
Treatment × Extreme Vote		0.006 (0.00044)	
Treatment × Voter Turnout			-0.022 (0.00127)
Day Time Trend	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes
Observations	154,616,084	147,194,110	147,194,110
R squared	0.113	0.110	0.110

Robust standard errors in parentheses clustered at the user level. In Panel A, the dependent variable is user level HHI in columns (1) and (2) while it is the number of clicks on Slot 4 related to politics. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period in columns (1) and (2) while in column (3) estimation is based on individuals we observe in the pre-experiment period as well. In Panel B, 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. The slight reduction in observations in (2) and (3) is due to unavailable demographic information.

been attributed to a lack of digital literacy with the new 'digital divide' being an 'algorithmic divide'. Individuals with extreme political views as well as a lack of political information is also associated with such behavior.[19] 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.

We carry out a survey to understand which observable characteristics in the main dataset are correlated with digital literacy. We find that individuals who do not read news online using desktop or laptop computers score significantly lower in our measure of digital literacy.[20]

[19] See `https://tinyurl.com/yybl4n58`.

[20] See the appendix and Table A.4 for details about the design and results of our representative survey.

Using this proxy, we show in column 1 of Table 4 that individuals that never accessed the news website through a desktop or laptop computers are more likely to increase their consumption share of politics when treated. Our other proxies are based on aggregate historical voting data. Individuals who reside in German states where there was a high share of votes to extreme political parties (right and left wing) in the last elections are more likely to increase their share of clicks on political stories click on the treatment. Additionally, regions with a higher voter turnout, a proxy for being more informed, are less likely to increase their click share for political news. Overall, with these results, we want to provide some grounding for assertions being made in the public discourse.

6 Conclusion

Our study with a large German news outlet using experimental variation within individual users across time, suggest that automated personalized recommendation can outperform human curation in terms of user engagement. However, we also highlight that this crucially depends on the amount of data available. Our results suggest that the human outperform algorithms when there is scare information on individual readers as well as limited data on fast developing news stories. During a time when there is a lot of discussion about which tasks will be automated, we find that human skills complement automated algorithms. We also find that initially, data related to individual reading behavior helps algorithmic effectiveness, but decreasing economic returns set in quickly and these returns taper off after a certain threshold. This has consequences for the recent policy debate related to privacy concerns and anti-competitive advantages data might bestow upon large firms. In particular, data might not provide a large strategic advantage over other firms and if data retention is to be limited due to privacy concerns, then it wouldn't significantly hurt the economic effectiveness of algorithmic recommendation.

We then show that there is an increase in concentration in the topics read by users when they are in the treatment group relative to when they are in the control group. This reduction in diversity of news consumption due to filter bubbles could have informational externalities in the public sphere. We also show using proxies of digital literacy and extreme political views that these individuals are more likely to be engaged by algorithmic recommendations. While our experiment is based on a subtle manipulation, we believe that these results are important in demonstrating behavioral patterns which are at the core of a recent public debate.

References

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

Athey, S., Mobius, M., and Pal, J. (2017). "The impact of news aggregators on internet news consumption." *Working Paper.*

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.

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.

Brynjolfsson, E., and Mitchell, T. (2017). "What can machine learning do? workforce implications." *Science*, *358*(6370), 1530–1534.

Calzada, J., and Gil, R. (2016). "What do news aggregators do? evidence from google news in spain and germany." *Working Paper.*

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

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

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

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

George, L. M., and Hogendorn, C. (2013). "Local news online: Aggregators, geo-targeting and the market for local news." *Working Paper.*

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

Lambrecht, A., and Tucker, C. (forthcoming). "Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads." *Management Science.*

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.

Mullainathan, S., and Spiess, J. (2017). "Machine learning: an applied econometric approach." *Journal of Economic Perspectives*, *31*(2), 87–106.

Oestreicher-Singer, G., and Sundararajan, A. (2012a). "Recommendation networks and the long tail of electronic commerce." *MIS Quarterly*, *36*(1).

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

Oh, H., Animesh, A., and Pinsonneault, A. (2016). "Free versus for-a-fee: The impact of a paywall on the pattern and effectiveness of word-of-mouth via social media." *Mis Quarterly*, *40*(1), 31–56.

Schaefer, M., Sapi, G., and Lorincz, S. (2018). "The Effect of Big Data on Recommendation Quality: The Example of Internet Search." *DIW Discussion Paper 1730.*

Shichor, Y. K., and Netzer, O. (2018). "Automating the b2b salesperson pricing decisions: Can machines replace humans and when?" *Working Paper.*

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`.

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.

A Appendix

Table A.1: Robustness: Logarithm of Clicks and Day Fixed Effects

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Slot 4	Other Slots	Total Hits	Slot 4	Other Slots	Total Hits
Treatment	0.0004	0.002	0.003	0.001	0.006	0.007
	(0.000)	(0.000)	(0.000)	(0.00004)	(0.00030)	(0.00030)
Day Time Trend	Yes	Yes	Yes	No	No	No
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	No	No	No	Yes	Yes	Yes
Observations	154,616,084	154,600,158	154,616,084	154616084	154616084	154616084
R squared	0.137	0.293	0.301	0.148	0.276	0.280

Robust standard errors in parentheses clustered at the user level. The dependent variable is log(1+number of clicks on Slot 4) in column (1), log(1+number of clicks on other slots) in column (2) and log(1+number of total clicks) column (3). The same variables in levels are in (4)-(6) but with day fixed effects instead of day time trend. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period.

Table A.2: Developments in German Parliament Coalition Formation

Breaking News Event	Dates
Merkel's Bavarian Allies Back In Play	18th-19th December 2017
Social Democrats Voted in Favor of Grand Coalition Talks	20th-23rd January 2018
Coalition Formation in Crisis with SPD Minister Resignation	4th-8th February 2018

Examples of these events and the lead up to such situations can be found here:
`https://www.politico.eu/article/spd-agrees-to-start-formal-coalition-talks-with-merkel/`
`https://www.politico.eu/article/martin-schulz-spd-i-wont-be-german-foreign-minister/`.

A.1 Supplementary Survey – Design and Results

Since the observational data in our main study does not include individual-level information that lets us directly classify a user's level of digital literacy, we conduct a supplementary survey. We access a panel of 500 German internet users through the crowdsourcing platform Clickworker – the German pendant to Amazon's MTurk. 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 to be representative of the German population (Wagner et al., 2007).[21] We construct an index of digital literacy using five survey questions (see Table A.3). 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.4 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. The size of estimated coefficient is about 15%. We cannot observe age, education and income in our main dataset, but we can observe whether a user never uses a laptop/desktop device. We therefore use this information as an individual-level proxy for digital literacy.

Table A.3: 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.4: 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	

White-robust standard errors in parentheses. The dependent variable is our digital literacy score.

[21] 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.