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

Jörg Claussen
LMU Munich[†]

Christian Peukert
CLSBE[‡]

Ananya Sen
MIT Sloan[§]

May 15th, 2019

Abstract

We run a large randomized field experiment with a major news outlet in Germany to quantify the economic returns to data and the informational externalities associated with algorithmic recommendation in the case of online news. We show that automated recommendation can outperform a human editor in terms of user engagement, though this crucially depends on the amount of training data. Limited data on individual behavior or on fast developing breaking news leads the editor to significantly outperform the algorithm. Additional data helps the algorithm beat the human editor but decreasing economic returns set in rapidly. Investigating the unintended consequences of personalized recommendation show that filter bubbles due to such recommendations reduce consumption diversity. Finally, in line with popular discourse, we demonstrate that users associated with lower levels of digital literacy and more extreme political views are more likely to engage with such algorithmic recommendations.

[*] We thank Marco Savioli and Lorien Sabatino as well as participants at the MIT CODE conference 2018, 16th Media Economics Workshop in Lecce, Munich Beliefs about Society and Politics Workshop and the Workshop on the Economics of Data and AI at the European Commission for helpful suggestions.

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

[‡] Catolica-Lisbon School of Business and Economics, christian.peukert@ucp.pt.

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

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) technologies are being utilized in a large number of industries to automate tasks which were previously being carried out by humans. The focus of automation till now has been on repetitive tasks performed by algorithms which involves minimal interpretation, creative and subjective judgment. It is unclear how humans would perform relative to algorithms in creative industries such as the news and media industry. 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 algorithm.[1] Whether humans making business decisions can outperform algorithms, potentially trained on a plethora of data, is an open question with important implications not only for firm strategy but also for competition and privacy policy. The potential existence of 'scale effects' in data used by algorithms has raised concerns among policy makers (for example the GDPR) regarding the anti-competitive advantage this could provide firms with access to rich consumer information.[2] Moreover,from a societal perspective, algorithmic recommendations might lead to a sub-optimal outcome if readers don't account for informational externalities of their own reading behavior, in particular if reader preferences are at odds with societal preferences. 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, data retention and informational externalities, we partner with a major German news outlet to carry out a field experiment.

[1] In fact, Apple News recently hired a sleuth of human curators instead of having automated 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.

[2] The regulatory environment in a number of countries restricts firms in how much specific information they can collect and store, has brought such issues at the forefront of the policy discourse both from a competition policy as well as a consumer privacy perspective. In fact, within the U.S., California has passed a Consumer Privacy Act which goes into effect in January 2020 exactly related to concerns about data retention and privacy.

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 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 allowing a recommendation algorithm, trained on fine grained individual level data, decide which of the N articles to be placed on a specific (fixed) slot $n = 4$.

In this setting, we first ask whether algorithmic recommendations can outperform a human editor in terms of user engagement (e.g. clicks)? Under what conditions can the human editor 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 marginal returns to data and how the precision 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 manipulate 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 and analyze how it might have been 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 anti-competitive issues by simply having access to individual level data. Additionally, if privacy concerns lead to limits on retention then this 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 on other slots related to similar topics while in the treatment condition relative to when they are in the control group. 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 and extreme political views are associated with a tendency to reduce consumption diversity in line with popular discourse.[3] These results speak directly to the bigger conversation at this time about the impact of filter bubbles and echo chambers because of algorithmic recommendations in the industry.

Our findings contribute to several streams of the literature. The first literature investigates 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.[4] 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

[3] https://www.niemanlab.org/reading/the-new-digital-divide-is-between-people-who-opt-out-of-algorithm

[4] 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”.

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.[5]

We also complement a few existing studies which look at the scale effects of data on precision of algorithmic predictions. 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 quality of search results do 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.[6] 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 in these studies.

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 reading consumption which goes beyond descriptive analysis and speaks to the issues raised by Gentzkow (2018).[7]

[5] There are other studies (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.

[6] 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”.

[7] 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.,

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 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 (eg. 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.[8] Our partner news outlet's coverage focuses on politics, finance and sports while also reporting on a variety of other topics. These different news beats have separate (human) editors. It is important to note that it is rare for few major news outlets in the world to experiment with algorithmic curation of their home page. 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 is exactly the same as seen by the control group. Moreover, the randomization is at the user-session level such that when the homepage is reloaded after a inactivity of thirty minutes, the randomization takes places again.[10]

The underlying model is a widely used supervised machine learning model (based on Bayesian networks) which is trained on fine grained data about the past reading behavior

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).

[8] There are several smaller news outlets which serve different types of readers by providing niche content.

[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. Even Google Search hardly has any degree of personalization except for based on location data (Gentzkow, 2018).

[10] 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 precision.

of each individual user.[11] The features used to train the algorithm employs 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.[12] 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 on other users' current reading behavior – the algorithm is not (necessarily) replicating the human editor's choice for slot 4. 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.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

[11] The algorithm implemented by the Data Science team is used in across industries as part of real time recommendation systems, for medical diagnosis, facial recognition and more.

[12] If the highest click probability article is already on a slot above (n=1,2 or 3), then the system chooses the next best.

if the user was in the control group (human curation). Theoretically, we should expect τ 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 which will determine there is a cannibalization or expansion effect overall. We include day fixed effects γ_τ to controls for events affecting all users, potentially through the news cycle. User fixed effects μ_i capture 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

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 turn to analyzing the impact of the treatment descriptively. Model free evidence in Table 2 does not suggest that the algorithm outperforms the human editor in terms of user clicks. We see that the number of clicks on slot 3 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

Table 1: Randomization Check

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

Standard errors in parentheses clustered at the user level. *** p<0.01, ** p<0.05, * p<0.1. The number of observations refers to individuals who we observe in the month before the experiment began. Column (3) measures the difference in means between the treatment and control group based on their pre-treatment characteristics and when they were first observed in the experiment.

suggests that personalized recommendation may lead to a positive attention spillover to the other slots and does not cannibalize their clicks.

Table 2: Model Free evidence

	(1) Control	(2) Treatment	(3) Difference((2)-(1))	(4) Magnitude	(5) Observations
Hits on Slot 4	0.0279	0.0276	-0.0003***	-1.1%	154,616,084
Hits on Other Slots	0.6902	0.6649	-0.0253***	-3.7%	154,616,084
Total Hits	0.7625	0.7438	-0.0187***	-2.4%	154,616,084

Standard errors in parentheses clustered at the user level. *** p<0.01, ** p<0.05, * p<0.1. Column (3) measures the difference in means between the treatment and control group. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period.

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 large number of one time visitors for which the ML model has no prior data. To explore this issue further, we turn to regression analysis.

Results from an OLS estimation of equation 2 in Table 3 paints a more nuanced picture. In column (1), when we have both individual and time fixed effects, we find that clicks that originate from slot 3 on the homepage increase by about 3% when it features a personalized recommendation, compared to the selection by the human editor. 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.

Table 3: Baseline and Scale Effects

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Slot=4	Slot≠4	Total	Slot=4	Total
Treatment	0.001***	0.006***	0.007***	-0.00772***	-0.0467***
	(0.000)	(0.000)	(0.000)	(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)
Time FE	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	0.179	0.141	0.389	0.181	0.141

Robust standard errors in parentheses clustered at the user level. *** p<0.01, ** p<0.05, * p<0.1. 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.

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 which measures the number of times user i has visited the website since November 2017 up until that session. In results reported in columns (4) and (5) of table 3, we find that the average user with a higher number of prior visits, and therefore more information available to the algorithm, increases clicks to articles on Slot 3 as well as overall clicks in a significant manner.

The above results, while illustrative, are still restrictive in analyzing the returns to data since we impose a linear structure on how engagement responds to prior data. 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:

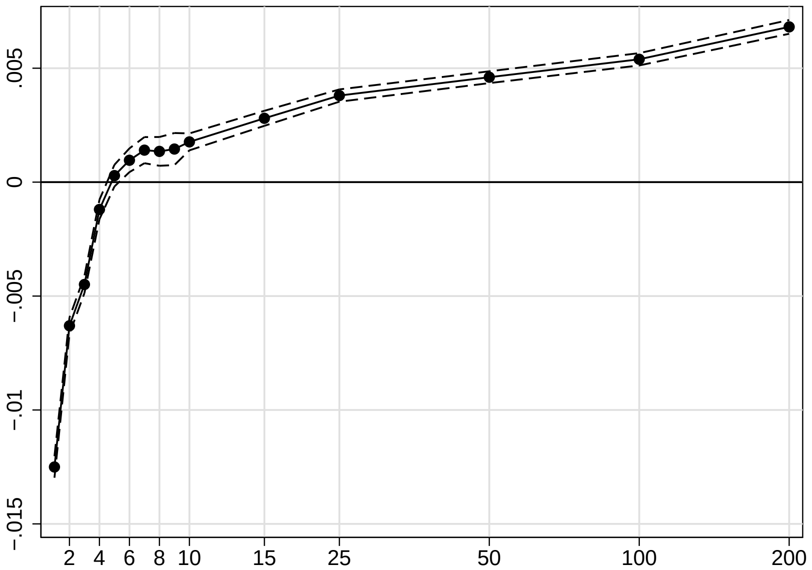
$$Clicks_{is} = \alpha + \delta_1 Treatment_{is} + \sum_q \delta_q (Treatment_{is} \times PriorVisits_q) + \gamma_\tau + \varepsilon_{is} \qquad (2)$$

$\forall q \in (0-2, 2-4, 4-6, 6-8, ..., 200-)$.

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 upto 10 visits per user then, the human has a comparative advantage. Around the threshold of ten visits, there is no 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, 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 that same statistical and economic level.

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

Figure 1: Decreasing Returns to data



related to the formation of the coalition between parties after the German parliamentary elections over a period of December 2017-February 2018. In column (1) of Table 4, 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 similar events in January and February related to the coalition talks to find very similar results in columns (2) and (3).[13] 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 is 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 returns set in quickly. The exact thresholds will, presumably, vary across

[13] 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/` and `https://www.politico.eu/article/martin-schulz-spd-i-wont-be-german-foreign-minister/`

Table 4: Breaking News and Algorithmic Performance

VARIABLES	(1) Slot=4	(2) Slot=4	(3) Slot=4	(4) Total	(5) Total	(6) Total
Treatment	0.001***	0.004***	0.001***	0.007***	0.000	0.007***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
Treatment × News (Dec)	-0.003***			-0.016***		
	(0.000)			(0.002)		
Treatment × News (Jan)		-0.008***			0.003	
		(0.000)			(0.002)	
Treatment × News (Feb)			-0.003***			-0.007***
			(0.000)			(0.002)
Time FE	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	0.182	0.22	0.184	0.411	0.42	0.22

Robust standard errors in parentheses clustered at the user level. *** p<0.01, ** p<0.05, * p<0.1. 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.

different contexts and algorithms. This result, though, is completely in line with the result of Chiou and Tucker (2017) where they show that a reduction in data retention to a threshold doesn't affect search engine performance. Our results also suggest that legislation as put forward by the European Commission might not erode the competitive edge of firms in a significant manner since adverse consequences on consumer engagement would be limited.

5 Information Externalities in Algorithmic Recommendations

In this section, we want to investigate the potential information externalities due to such 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. 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.

Table 5: Algorithms and Information Externalities

	(1)	(2)	(3)	(4)
VARIABLES	User HHI	User HHI	(Post) Politics	(Post) Nat. Pol.
Treatment	0.012*** (0.000)	0.008*** (0.000)	0.002*** (0.001)	0.003*** (0.000)
Visits		-0.304*** (0.000)		
Treatment x (Pre) Politics			0.003*** (0.000)	
Treatment x (Pre) Nat. Pol.				0.004*** (0.000)
Observations	12,803,057	12,803,057	63,706,396	63,706,396
R squared	0.182	0.22	0.184	0.411

Robust standard errors in parentheses clustered at the user level. *** p<0.01, ** p<0.05, * p<0.1. 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 in columns (1) and (2). In columns (3) and (4), estimation is based on individuals we observe in the pre-experiment period as well.

The results in Table 5 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 if include controls beyond user fixed effects, such as the number of times the user has been on the website (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. 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 with this effect being driven by consumption of national political news.

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'. Individuals with extreme political views as well as a lack of political information is also associated with such behavior.[14]

Table 6: Consumption Diversity and Reader Characteristics

	(1)	(2)	(3)
VARIABLES	Share Politics (Slot4)	Share Politics (Slot4)	Share Politics (Slot4)
Treatment	0.006***	0.004***	0.022***
	(0.000)	(0.000)	(0.001)
Treatment × Apple User	-0.001***		
	(0.000)		
Treatment × Extreme Vote		0.006***	
		(0.000)	
Treatment × Voter Turnout			-0.022***
			(0.001)
Time FE	Yes	Yes	Yes
Observations	154,616,084	147,194,110	147,194,110
R squared	XX	XX	XX

Robust standard errors in parentheses clustered at the user level. *** p<0.01, ** p<0.05, * p<0.1. 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.

We test for these hypotheses by using proxies for such characteristics. Analyzing these heterogeneous treatment effects can be an informative exercise to provide evidence for debates in the popular press. Proxying higher digital literacy by whether the individual

[14] See `https://www.washingtonpost.com/news/the-intersect/wp/2018/05/25/how-googling-it-can-send-conservatives-down-secret-rabbit-holes-of-alternative-facts/?utm_term=.ebdd84726d29`

uses an Apple device (mobile or desktop), we find that such readers are less likely to read politics since they reduce their politics click share on Slot 4 when they are in the treatment group (column (1), Table 6). 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 more 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 prediction a lot but decreasing economic returns set in quickly and these returns taper off after a certain threshold. This has consequences for the existing 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 quality 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 being debated in the popular press.

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.

Blankespoor, E., deHaan, E., and Zhu, C. (2018). "Capital market effects of media synthesis and dissemination: Evidence from robo-journalism." *Review of Accounting Studies*, *23*(1), 1–36.

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.*

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*.

Supplementary Appendix