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

Jörg Claussen
LMU Munich and CBS[†]

Christian Peukert
CLSBE and ETH Zurich[‡]

Ananya Sen
Carnegie Mellon University[§]

November 3, 2019

Abstract

We run a field experiment to quantify the economic returns to data and informational externalities associated with algorithmic recommendation in the context of online news. Our results show that personalized recommendation can outperform human curation in terms of user engagement, though this crucially depends on the amount of personal 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 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, CIST, Paris Seminar on the Economics of Digitization, and Zurich Media and Digitization Workshop 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) and Copenhagen Business School, j.claussen@lmu.de.

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

[§] Carnegie Mellon University, ananyase@andrew.cmu.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 (Brynjolfsson and Mitchell, 2017). Agrawal et al. (2018) hypothesize that "the future's most valuable skills will be those that are complementary to prediction in other words, those related to judgement". In line with this logic, it is unclear how humans would perform relative to algorithms in 'creative' industries, such as the news and media industry, since editorial decisions necessarily involve subjective judgments about 'newsworthiness' of stories.[1] Whether humans making business decisions can outperform algorithms, potentially trained on a plethora of data, has important implications for competition and privacy policy as well. The rise in market power of online platforms has caught the attention of policy makers, who seem to believe that 'scale effects' of data can be a significant source of anti-competitive behavior.[2] Empirical evidence for assertions on either perspective, though, is limited. Relatedly, in the case of online news, 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).

To explore these interrelated issues of algorithms, economic returns to data, and informational externalities, we partner with a major news outlet in Germany to carry out a field experiment. The homepage of the news outlet's website is 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 home-

[1] 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, Facebook has recently hired a slew of human editors to curate news stories for a new venture after disbanding their algorithmic "Trending News" feature. See `https://www.nytimes.com/2019/08/20/technology/facebook-news-humans.html` for more details on this venture. Moreover, Apple News has also recently hired a number of journalists 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.

[2] The regulatory environment in a number of regions (e.g. GDPR, California Consumer Privacy Act) is putting restrictions on what firms can do with user data, bringing these issues to the forefront of the debate on competition policy and consumer protection. 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.

page, she is randomly assigned to a control or treatment condition. If a user is assigned to the control condition, then all the articles she observes on the homepage are the ones curated by the human editor. In the treatment condition, the homepage is customized to show recommendations from a recommendation algorithm, trained on browsing information from that individual user as well as browsing behavior of other users.[3] In this setting, we first ask whether algorithmic recommendations can outperform a human editor in terms of user engagement (clicks) and under which circumstances 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.[4] More generally, we investigate the economic returns to data and how the effectiveness of the algorithmic recommendation improves with more personal data, relative to the human editor. We analyze how the treatment effect varies as more individual-level as well as aggregate browsing data is used to make recommendations. Finally, we look into the potential information externalities. We test whether consumption diversity across news categories is impacted by personalized algorithmic recommendations, and analyze which user characteristics are associated with changing behavior. Since the algorithm makes recommendations based both on data from the individual user as well as from other users, the direction of change in consumption diversity is ambiguous from a theoretical perspective.

We find two broad sets of results. First, we show that the algorithmic recommendations on average receive more clicks than the human-curated version of the homepage. Controlling for user-specific unobserved variation in a fixed effects model, we show that users in the treatment group increase clicks to the treatment slot by about 4% and overall clicks by about 1%. More generally, we find that in cases where the algorithm has limited individual level data, the automated recommendations lead to less clicks than the human-curated control condition. Only users about which the algorithm has browsing information from more than five visits make more clicks when in the treatment condition. While data helps algorithmic performance, we show that there are diminishing *economic returns* to data which set in rapidly with data from an additional visit leading to smaller increases in click rates. Additionally, we show that the

[3] Based on its estimate of the category that maximizes the probability that the individual will click, the algorithm decides which of the N articles to be placed on a specific (fixed) slot $n = 4$. Hence, the algorithm uses 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.

[4] See for example the discussion in `https://www.theguardian.com/commentisfree/2007/dec/28/theeditorascurator`.

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 identifying the taste of the average reader when an algorithm has limited data implying that a combination of the human and algorithmic editor might provide the biggest payoff to the firm.

Second, we find that algorithmic recommendation reduces the consumption diversity by users when they are in the treatment group 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, when being in the treatment group. Additionally, we show that proxies of digital literacy[5] and extreme political views are associated with a tendency to reduce consumption diversity in line with popular discourse.[6]

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 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 exogenous variation. Moreover, our setting is rich enough for us to provide a full characterization of the economic returns to data, which could help reconcile the effects found in the 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 and asymptotic theory.

[5] 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.4 and A.5 in the Supplementary Appendix for details.

[6] See for example Susarla (2019).

In this paper, by focusing on economic returns, we are also attempting to map data into reader preferences and firm revenue using experimental variation, while accounting for time variant user level unobserved heterogenity and prior user experience. A priori, policy makers have been concerned about discontinuities, threshold effects and increasing returns at different intervals as we map additional data for the algorithm into economic outcomes.[7] Our results indicate that there might be limited strategic advantages for a firm by simply having access to 'big data' and if privacy concerns lead to limits on retention then this should not have too big an impact on algorithmic performance.

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.[8] 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 humans 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.[9]

Analyzing the externalities of personalized news recommendation also contributes to the literature on the role of digitization 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 consumption diversity, in some cases 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 debate about how AI can affect the efficiency with which the market matches news content to consumers outlined

[7] See for example a report related to a hearing at the Federal Trade Commission `https://drive.google.com/open?id=1Kh4tzncv15W9fmuK1wNMRLSs7Z0gmqjT` and by the European Commission `https://ec.europa.eu/jrc/communities/sites/jrccties/files/dewp_201809_data_and_ai_181218.pdf`. A recent paper by Brynjolfsson et al. (2018) shows that a marginal increase in precision of machine translation, from 85% to 90% accuracy, on eBay can increase online international transactions by upto 20%.

[8] 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".

[9] 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.

in Gentzkow (2018).[10]

2 Background and Experimental Setting

2.1 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, has recently started experimenting with personalization of an individual reader's newsfeed only based on geographical location.[11]

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. The fact that news editors care about increasing advertising revenue explicitly and choose news stories based on those calculations has been highlighted extensively in the media economics literature (for example, see Gentzkow and Shapiro, 2010 and Sen and Yildirim, 2015). Moreover, even if editors only care about 'impact' driven stories, evidence shows that 'important' stories are highly correlated with greater audience reach as measured by clicks (Sen and Yildirim, 2015). Hence, focusing on clicks as the main outcome variable of interest would be a feasible way to capture various objectives in a reduced form manner.[12]

[10] 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).

[11] 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.

[12] 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

2.2 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. 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 layout of the website's homepage is such that the articles appear one below the other and not side by side. Figure 1 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, the order of which is decided by the human editor. The algorithm chooses from the same underlying pool of articles but, for example, "bumps up" the article in slot 6 to slot 4. If the article chosen by the algorithm is already on a slot above (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 while maintaining their general ordinal rank.[13] The experiment was carried out from December 2017 to May 2018 and included all visitors, across 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 when the user is inactive for thirty minutes and/or reloads the homepage, the randomization takes places again. This level of randomization, compared to others, provides us with sufficient statistical power to utilize meaningful variation within users. More importantly, employing user fixed effects allows us to separate the impact of time invariant unobserved user heterogeneity and user experience from the effect of the amount of user-specific data on algorithmic effectiveness which is the core issue in this paper.[14] 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 3-5% 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 economic returns to data at the individual level and associated changes in consumption diversity.

The algorithm implemented uses the method developed by Google engineers for Google News,

more detail below.

[13] It is important to note that there are about 80 articles on the homepage at any one point in time which provides a broad pool for the algorithm to choose from. Moreover, no algorithm is used for producing any content. Journalists employed by the news outlet produce all the available content.

[14] 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 4.3.

the details of which were published in Liu et al. (2010). The objective of the algorithm was to maximize the clicks on the recommended article. The recommender systems combines personal data for predicting individual interest based on past browsing behavior, and social data to assess current interests of other users. Using past individual clicking behavior along with current news trend information ensures that while catering to the preferences of the individual reader, the model does not miss out on current news events. A user is identified based on a unique cookie ID. 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 clicking likelihood 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.[15] If a user has no prior reading behavior, then the system assigns a recommendation that is based on current news trends which are in turn driven by features across other users' current reading behavior. While the model predictions will be based on the current news trend when there is zero personal data available, its predictions will increasingly be based on users personal data as the browsing behavior by an individual increases.[16] The algorithm is not (necessarily) replicating the human editor's choice for slot 4.

3 Empirical Framework

Our baseline specification links a reader's engagement on the website to treatment status:

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

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*), other slots on the homepage (*Slot≠4*), or the number of total clicks in a given session, which can include clicks on articles

[15] 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 sevral hundred topics. A reader's change in preferences across topics are accounted for on a daily basis like in Liu et al. (2010).

[16] See Section A of the Supplementary Appendix for more on the technical details of the algorithm sketched out in Liu et al. (2010) and used as a baseline framework by the Data Science team.

not highlighted on the homepage.[17] 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 and total clicks in a session 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 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.

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 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. As can be seen, the sample is well balanced across all the observables, indicating that our randomization has worked in the desired manner.[18]

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 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.[19] 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 estimate can hide important

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

[18] The summary statistics of different variables of interest are provided in Table 2 in the appendix. We use these baseline measures to discuss magnitudes of our estimates.

[19] Effect sizes are reported as relative to the baseline, i.e. the sample average. See Table 2 for summary statistics. In the case of column (1), we compare 0.0003 to 0.027 which is the mean number of clicks to Slot 4. We calculate other estimates similarly.

heterogeneity. For example, our sample includes a number of visitors who arrive on the website only a few times, for whom the ML model has limited prior data.[20] To explore this issue further, we turn to regression analysis with fixed effects that can capture unobserved heterogeneity across users and time.

Results from an OLS estimation of equation 1 in columns (2)-(6) of Table 3 paints a more nuanced picture. In column (2), 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. This specification accounts for differences in reading behavior across users, reducing the bias in the estimates from a simple OLS model. In column (3), 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 (4), 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 Economic Returns to Data

4.2.1 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, as we show that the estimate of the treatment effect increases once we control for user-specific heterogeneity that also includes a user's number of visits to the homepage. Next, we go on to explore heterogeneity in the treatment effect more directly by testing for scale effects of the algorithm's access to personal data as measured by the number of a user's prior visits to the homepage. 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 has visited the website since December 2017 up to that session. In results reported in columns (5) and (6) of Table 3, 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.

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

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:

$$Clicks_{is}^{k} = \alpha + \delta_1 Treatment_{is} + \sum_{q} \delta_q (Treatment_{is} \times PriorVisits_q) + \lambda_q PriorVisits_q + \gamma_\tau + \varepsilon_{is} \tag{2}$$

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

The results in Figure 2 provide an insightful overview to the the average returns of data to the algorithm relative to the human editor. It plots how much more readers click on slot 4 when in the treatment group relative to the control at different levels of past visits, controlling for user fixed effects and the average clicks at that number of past visits across users. Using our experimental setup, user fixed effects and controls for the average clicks at each level of past visit, we can quantify the returns to data in clean manner. 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 10 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, 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.[21]

It is 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 individual level data can

[21] Algorithmic performance remains at similar economic levels even when we extend the series with finer intervals. As expected given the experimental setup, the general finding is independent on user experience. Figure A.1 in the appendix shows that there are also decreasing returns to data when we only look at users that we can observe before and during the experiment (left-hand panel), and when we only look at users that we can only observe during the experiment (right-hand panel).

help firms gain a competitive advantage but diminishing economic returns set in quickly. Hence, the competitive advantage for firms from employing data in algorithms might be limited.[22] 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 existing studies such as that by Chiou and Tucker (2017). 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 algorthimic recommendations might have been on the 'flat' part of the curve in Figure 2 where additional data does not add much value to the recommendations.

4.2.2 Breaking News, Editorial Judgment and Algorithmic Predictions

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, who have domain knowledge and expertise, are better at forecasting the 'newsworthiness' of a big developing story. More generally, it is important to note that being 'first to market' with a big news story 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 breaking the big story in a matter of minutes (Cagé et al., 2017). 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 3 illustrates these spikes in public interest on particular days with respect to politics and sports in Google search data.

In Table 4, 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 breaking news events.

[22] On the flip side, from a privacy policy perspective, this might suggest that legislation put forward by various institutions, including the European Commission (GDPR) 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.

We repeat this exercise for big sport events and clicks to sports articles on the treatment slot to find very similar results in columns (3) and (4).[23]

Overall, we find that 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. Moreover, the human outperforms the algorithm when there are fast developing breaking news stories which the human is quicker to identify than the ML model. This is true even though the algorithm used in our setting was designed to not only include individual-level information, but also general news interest as measured by engagement of other users. We believe that such differences between the human and the algorithm due to limited data are fundamental and are likely to persist in the future. The 'cold start' problem of algorithms due to limited data at the beginning of a reader's journey might provide an edge to humans in other contexts as well. Moreover, it is the domain knowledge and expertise of an editor which allows her to identify the potential of a fast paced breaking news event while an algorithm plays catch up after that particular news story has gathered a sufficient number of clicks. 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.

4.3 Validation, Alternative Variation and Robustness

Our baseline results outlined above provide a broad yet nuanced picture of the returns to data in our context of the editor and the algorithm. While our estimates are based on fine-grained data from a randomized experiment, we want ensure that our results rule out other potential explanations which might impact the magnitudes.

4.3.1 The New Year Bug: A Natural Experiment

First, we assess the validity of the algorithm performance becoming better with additional data by exploiting a natural experiment within our sample period. In particular, we were informed by the data science team at our partner news outlet that they had identified a bug in their code which impaired the ability of the algorithm to update its recommendations for the first six days of January 2018. 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. This means that the recommendations were based on less recent data both in terms of the general news cycle but

[23] Our results are robust to alternative time thresholds for these events.

also personal behavioral data captured by user specific clicks. Moreover, as the days went on, the data utilized would become even less recent and relevant and should in fact, make the recommendations worse. Hence, an individual who came on to the news outlet's website on the $1st$ of January would see more 'relevant' content than if she or any other user came on the $6th$ of January since the recommendations by the algorithm would now be more outdated. This provides exogenous variation in data quantity which we can exploit for another exercise to provide causal interpretation to our results.

We focus on data from December 2017 and January 2018 of our sample and code 'New Year Bug' equal to one if it was one of the six days 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 individual was observed during the coding error days. Column (1) of Table 5 shows that conditional on the fixed effects, clicks on slot 4 reduced in a significant manner when a user was in the treatment relative to the control during the six days of when the coding bug went unnoticed (the interaction term). In column (2), instead of a dummy variable, 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 as each day went by, the data utilized became less recent and less personalized leading to fewer clicks on our treatment slot. Additionally, as before, we find that the treatment effect is positive and statistically significant. Overall, this gives us further confidence that our baseline results are indeed capturing the returns to data.

4.3.2 Alternative Sources of Variation and Robustness

Next, we want to analyze how our baseline estimates might vary across different dimensions of heterogeneity such as regular vs. irregular users and analyzing users who were only observed in our pre-treatment phase. One would hypothesize that both these groups of users provide the algorithm with more data and hence we should see higher than average effect in terms of hits on the treatment slot. Columns (3) and (4) of Table 5 point exactly in that direction with positive and statistically significant estimates of 0.003 (regular users who visited the website more than 10 times during the sample period) and 0.005 (pre-experiment users). These estimates are higher than those found in the overall sample exactly because we observe them for longer which allows improvements in algorithmic performance.

Next, we carry out a few checks on 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 A.2 in the Appendix report the results using within user-week and user-day variation to find results which are qualitatively and quantitatively similar to our baseline estimates. Column (3) and (4) use user-hour and user-hour of the date variation to find remarkably similar results. This gives us confidence that our results are not being driven by differences in reading behavior due to changes in the news cycle or by different preferences or strategies adopted by human editors.

In columns (1)-(4) of Table A.1, we look at alternative functional forms for the dependent variable with the logarithm of clicks in (1)-(2) and the probability of any click in (3)-(4) to find qualitatively similar results. Finally, in columns (1)-(4) of Table A.3, we use some more sources of variation to further assess the robustness of our results. In column (1), we use the first session of the morning for a user (6 am to 12 pm) while column (2) uses the first session of the afternoon (12pm to 6pm) to find similar results as in the baseline. In column (3) we restrict the data to look at only the first session of every hour for a user, and finally, column (4) utilizes data from only the first session of every day for a user. For each of these specifications in (1)-(4), we find that the results are qualitatively and quantitatively in line with our baseline estimates in column (5) of Table 3. This gives us confidence that our results are indeed being driven by the treatment and not due to abnormal reading patterns and dynamics of online users.

4.4 Revenue Implications for Automating Editorial Curation

Our analysis above 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 the economic size of our estimates into context with a simple back of the envelope calculation. In particular, we want to assess how much revenue a news outlet might be able to generate if they automated the curation process completely relative to the costs of implementing such a system which would include, in particular, 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.[24] 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, which is about 22% of the wage, comes to a total of $7,200 (approx.) which can be considered as the cost of implementing such an automated system.[25] 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 profitable.[26] 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.[27]

The aim of this exercise is to demonstrate how to use these estimates to evaluate alternative scenarios for news outlets with different audience sizes (local vs. national), number of ads on a page, and algorithmic performance.[28] A caveat, of course, is that these estimates are partial equilibrium since its not necessarily the case that a data scientist has to be assigned full time to this task and moreover, such a change could also free up the curating human editors to carry out other tasks. Further, we of course abstract from general equilibrium effects that might arise with competition across outlets.

[24] 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.

[25] 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.

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

[27] See `https://www.similarweb.com/website/bostonglobe.com` for aggregate statistics on this.

[28] In fact, as just another scenario, we would find similar results if there's only one ad per page but there were spillovers across clicks on articles as we found in our setting which would make the estimate of 9% from column (1) of Table 3 more suitable to use.

5 Information Externalities in Algorithmic Recommendations

News is a special 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.[29] 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 or categories of articles over the experimental period. We are interested in not only the impact of the clicks on the algorithmically recommended article but also the follow-on clicks 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 article categories (such as politics, sports, finance, etc.) coming both from clicks on slot 4 as well as the resulting spillovers onto other slots.

The direction of change in consumption diversity, if any, is ambiguous since the algorithm is trained on past individual level reading behavior as well as current news trends based on the browsing 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.

We use the Hirschman-Herfindahl Index (HHI) measure of consumption shares across different topics or categories at the individual user 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 takes into account 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 increase. We map this definition into our context of reading behavior 'concentration' across topics. Since the article categories through our sample period remain unchanged, an increase

[29] 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.

in HHI would be the result of an increase in the disparity of the relative distribution of clicks across topics from a reader implying an increase in concentration across topics.

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 over the course of 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 6 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 on to other slots as well. The magnitudes imply that there was an increase in user level HHI by 5% for slot 4 and by 0.5% in terms of spillovers to other slots.[30] To dig deeper, we use pre-experimental browsing behavior from November 2017 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 been attributed to a lack of digital literacy with the new 'digital divide' being an 'algorithmic divide'.[31] Individuals with extreme political views as well as a lack of political information are also associated with such behavior.[32] 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 the data in our sample does not include individual-level co-variates that let 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 crowd-sourcing platform Clickworker – the German pendant to Amazon's MTurk.[33] We construct an

[30] 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.

[31] See the discussion in Susarla (2019).

[32] See for example the following article in the Washington Post: https://tinyurl.com/yybl4n58.

[33] 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). The average age of an internet user in SOEP

index of digital literacy using five survey questions (see Table A.4). 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.5 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.[34]

Results in column (1) of Table 7 suggest that individuals that did not access the news website through a desktop or laptop computers are more likely to increase their consumption share of politics when treated. For proxies of extreme political views, we use state level voting by looking at the share of votes going to extreme right and left parties in the last elections. Following Larcinese (2007), we use voter turnout in the last election as a proxy for political knowledge which is also aggregated at the level of the state. Results in column (2) and (3) of Table 7 suggest that 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 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. Overall, with these results, we aim to provide some grounding for assertions being made in the public discourse.

6 Discussion

The results presented in this paper are based on one field experiment with a large German news outlet. Hence, one question that can arise is whether we should expect the patterns we have uncovered in the case of this experiment to be repeated in other contexts. We believe this to be the case. First, it is evident from the debate in the news industry about how even the more traditional news aggregators such as Apple News and Facebook are reverting back to having human editors in the mix. This gives us confidence that such developments provide validity on the broad question we attempt to address. Moreover, while there is limited anecdotal evidence that mainstream news outlets algorithmically curate their homepage, one would expect that such experimentation would begin in the near future. Additionally, the algorithm utilized in our field

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.

[34] As some external validity, Qian et al. (2019) find that individuals using only mobile search for online shoping are based in more economically dis-advantaged regions.

experiment has been implemented at Google News and is based on a highly cited academic paper. While different algorithms can be used in other settings, a widely used algorithm especially by major players such as Google News provides some level of credence to our results.

Next, even though our experiment involves a subtle treatment it allows us to characterize a rich set of results. We first highlight how humans and algorithms could be complements rather than substitutes. We believe that this stems from a core difference between humans and algorithms in their ability to deal with limited data, on a variety of dimensions, which would extend into other domains. This intuition could apply in the context of the launch of new products as well hiring new employees, or more generally where 'subjective judgment' (Agrawal et al., 2018) or domain expertise and knowledge are valuable due to limited training data being available.

Moreover, our ability to track an individual and see how behavior varies with more data allows us to reconcile some of the mixed findings in this literature quantifying the value of data. Our findings using experimental variation have relevance in explaining the null results in the context of search engine recommendations by Chiou and Tucker (2017) and provide qualitatively similar results as in the case of product demand forecasting by Bajari et al. (2018), which is more descriptive in nature since they use observational data without any quasi-experimental variation.

Finally, our results on the diversity of news consumption also relates to how algorithmic recommendations might affect product consumption in other contexts. This finding of reduction in consumption diversity is similar to that of Hosanagar et al. (2014) who look at collaborative filtering recommendations in online shopping behavior. 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.

7 Conclusion

Our study with a large German news outlet using experimental variation within individual users across time, suggests that automated personalized recommendation can outperform human curation in terms of user engagement. While this is of course a rather unfair match, because human curation can never be individualized at scale, we also highlight that the relative better algorithmic performance crucially depends on the amount of available data. Our results suggest that the human curation outperforms the algorithm 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. The particular returns to data that we identify can be informative more generally. Our findings contribute to the recent policy debate related to privacy concerns and competitive advantages data might bestow upon large firms. In particular, if data retention is to be limited due to privacy concerns, then our results suggest that it would not 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.

We conclude by highlighting some limitations of our study which should create opportunities for future work. 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. While our findings could apply in different online settings, this literature is still nascent, and hence we need more studies to provide a richer picture related to the returns to data. While we've tried to differentiate between social and individual data, there is a lot of scope to dig deeper into the importance of different types of data in different settings. Additionally, it would be important to dig deeper into which kinds of users are likely to engage more with algorithms and why that might be happening. Our exploratory 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' to garner clicks.[35]

[35] See https://www.nytimes.com/2019/03/29/technology/youtube-online-extremism.html as an example.

References

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

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

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.

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.

Brynjolfsson, E., Hui, X., and Liu, M. (2018). "Does machine translation affect international trade? evidence from a large digital platform."

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

Cagé, J., Hervé, N., and Viaud, M.-L. (2017). "The production of information in an online world: Is copy right?"

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.*

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.

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.*

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

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.*

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

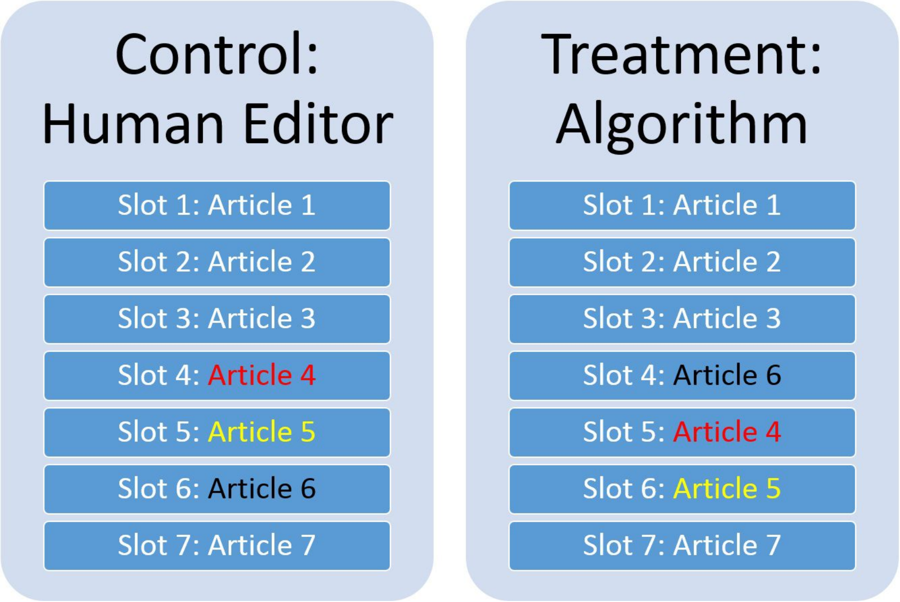
Shichor, Y. K., and Netzer, O. (2018). “Automating the b2b salesperson pricing decisions: Can machines replace humans and when?” *Working Paper.*

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.

Figures and Tables

Figure 1: 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)	(2)	(3)	(4)	(5)
VARIABLES	Control	Treatment	Difference((2)-(1))	Std. Error	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

Column (3) measures 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. We normalize these numbers using a scaling factor.

Table 2: Summary Statistics

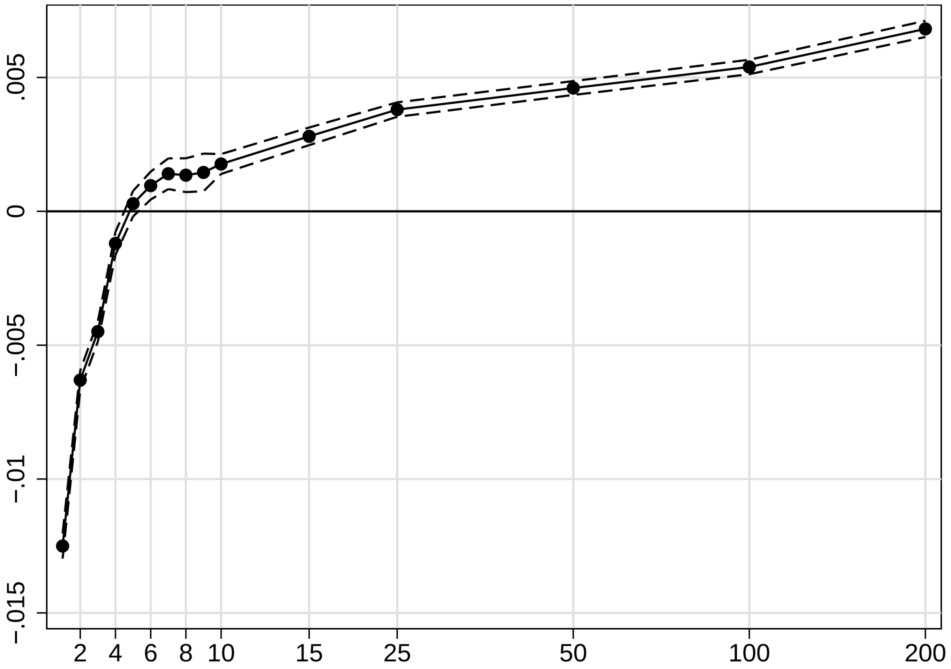
VARIABLES	Obs.	Mean	Std. Dev.	Min.	Max.
Hits on Slot 4	164,982,192	0.0275075	.1845144	0	110
Hits on Other Slots	164,982,192	0.7279069	1.35175	0	2955
Total Hits	164,982,192	0.7553178	1.365205	0	2955
Politics Share (Pre-Treatment Users)	63,706,396	1.420107	2.044066	0	312.5
Hits on Slot 4 (Pre-Treatment Users)	63,706,396	0.0295634	0.1890511	0	110
Hits on Slot 4 (New Year Bug)	10,366,108	0.0242998	0.1659694	0	54

Table 3: Baseline and Scale Effects

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Slot=4	Slot=4	Slot≠4	Total	Slot=4	Total
Treatment	-0.0003***	0.001***	0.006***	0.007***	-0.00772***	-0.0467***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Treatment × Prior Visits					0.00287***	0.0190***
					(0.000)	(0.000)
Day FE	No	Yes	Yes	Yes	Yes	Yes
Individual FE	No	Yes	Yes	Yes	Yes	Yes
Observations	154,616,084	143,695,807	143,695,807	143,695,807	143,695,807	143,695,807
R-squared	0.00004	0.093	0.204	0.208	0.093	0.208

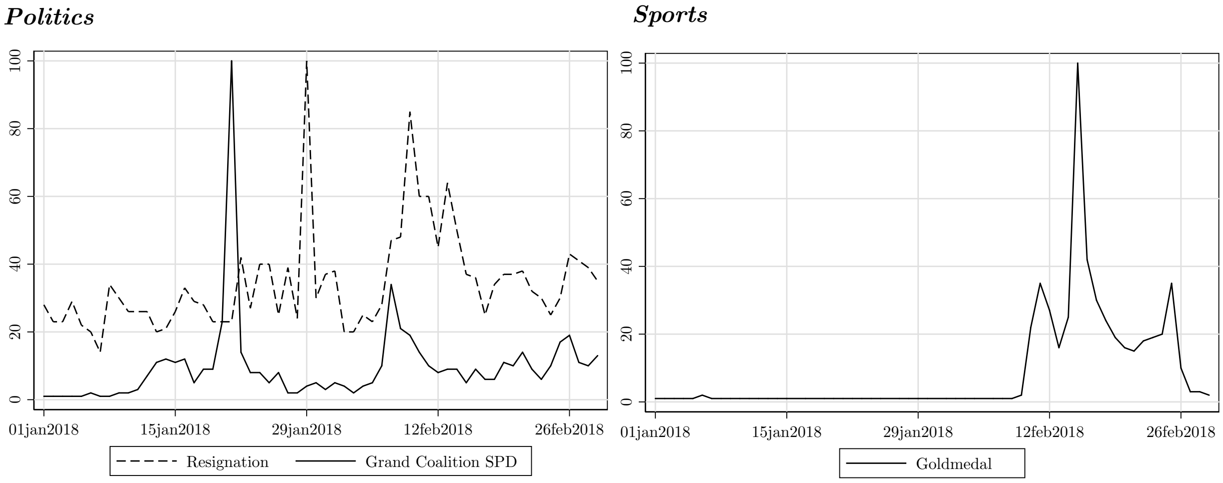
The dependent variable is the number of clicks on Slot 4 in columns (1), (2) and (5), total clicks in the session in (4) and (6) and clicks on other slots in column (3). *Prior Visits* is the user's log number of prior visits to the homepage since the beginning of the experiment. The unit of observation is user-session. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Figure 2: Decreasing Returns to Data



The figure plots the coefficients δ_q along with 99% 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.

Figure 3: Breaking 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 4: Breaking News and Algorithmic Performance

VARIABLES	(1) Slot=4	(2) Slot=4	(3) Slot=4
Treatment	0.009***	0.008***	0.0004***
	(0.00005)	(0.00004)	(0.00007)
Treatment x GrandColSPD	-0.001***		
	(0.00020)		
Treatment x Resignation		-0.005***	
		(0.00015)	
Treatment x Goldmedal			-0.008***
			(0.00020)
Day FE	No	No	No
Individual FE	Yes	Yes	Yes
Observations	38,286,023	63,887,513	24,745,728
R-squared	0.112	0.107	0.127

The dependent variable is the number of clicks on politics articles on slot 4 in columns (1)-(2) and number of clicks on sports articles on slot 4 in (3). The unit of observation is user-session. The number of observations includes all individuals observed during the experimental period, in columns (1) and (3) during January 2018, and in columns (2) and (3) during January and February 2018. News events are defined as indicating the days of the respective spikes in Figure 3. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 5: New Year Bug, Heterogeneity and Algorithmic Performance

	(New Year Bug)	(New Year Bug)	(Regular Users)	(Pre-Treatment Users)
	(1)	(2)	(3)	(4)
VARIABLES	Slot=4	Slot=4	Slot=4	Slot=4
Treatment	0.002***	0.002***	0.003***	0.005***
	(0.00006)	(0.00006)	(0.00004)	(0.00006)
Treatment × New Year Bug	-0.005***			
	(0.00013)			
Treatment × New Year Bug Day Trend		-0.001***		
		(0.00003)		
Day FE	Yes	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes	Yes
Observations	74,168,858	74,168,858	118,780,993	61,194,540
R-squared	0.102	0.102	0.042	0.024

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 at least ten times while column (4) looks at only those who were observed in our pre-treatment sample. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 6: Information Externalities 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*** (0.001)	0.025*** (0.001)
Treatment x (Pre) Politics			0.003*** (0.0004)	0.003*** (0.001)
Individual FE	Yes	Yes	Yes	Yes
Observations	2,446,445	23,426,694	63,403,203	63,403,203
R-squared	0.120	0.640	0.027	0.148

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) and (4) estimation is based only on individuals we observe in the pre-experiment period as well as during the experiment. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table 7: Information Externalities: Consumption Diversity and Reader Characteristics

VARIABLES	(1) Share Politics (Slot4)	(2) Share Politics (Slot4)	(3) Share Politics (Slot4)
Treatment	0.005*** (0.00002)	0.004*** (0.00009)	0.022*** (0.00090)
Treatment × No Desktop/Laptop	0.002*** (0.00008)		
Treatment × Extreme Vote		0.006*** (0.00041)	
Treatment × Voter Turnout			-0.022*** (0.00118)
Day FE	Yes	Yes	Yes
Individual FE	Yes	Yes	Yes
Observations	143,695,807	137,424,038	137,424,038
R-squared	0.088	0.086	0.086

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

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

$$\begin{aligned} p_{i,j}^t(cat = c_j | click_i) &= interest_{i,j}^t(cat = c_j) \times G(interest_j^\gamma(cat = c_j)) \\ &= \frac{p_j^\gamma(cat = c_j) \times \left(\sum_{\tau=1}^{t-1} \left(N_i^\tau \times interest_{i,j}^\tau(cat = c_j) + G\right)\right.}{\sum_{\tau=1}^{t-1} N_i^\tau + G} \end{aligned} \tag{3}$$

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 3 it becomes clear that the model's predictions will be entirely based on the current news trend when there is zero personal data available, and the model's predictions will increasingly based on user i's personal data $\sum N_i$ grows larger than G.

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.

Table A.1: Robustness: Logarithm and Probability of Clicks

	Log(Clicks)	Log(Clicks)	Prob(Click)	Prob(Click)
	(1)	(2)	(3)	(4)
VARIABLES	Slot=4	Total	Slot=4	Total
Treatment	0.0004***	0.003***	0.0004***	0.003***
	(0.000)	(0.0001)	(0.0000)	(0.0001)
Individual FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Observations	143,695,807	143,695,807	143,695,807	143,695,807
R-squared	0.137	0.301	0.135	0.265

The dependent variable is log(1+number of clicks on Slot 4) in column (1), log(1+number of total clicks) in column (2) while it is the probability of any click on Slot 4 in column (3) and probability of any click in the session in column (4). The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period for columns (1)-(4). Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.2: Alternative Baseline Specification with Finer Fixed Effects

	(1)	(2)	(3)	(4)
VARIABLES	Slot=4	Slot=4	Slot=4	Slot=4
Treatment	0.001***	0.002***	0.002***	0.002***
	(0.00004)	(0.00005)	(0.00004)	(0.00005)
Day FE	Yes	Yes	Yes	Yes
Fixed FE	User-Week	User-Day	User-Hour	User-Hour of Day
Observations	144,926,111	145,275,590	143,453,411	141,598,937
R-squared	0.116	0.151	0.129	0.171

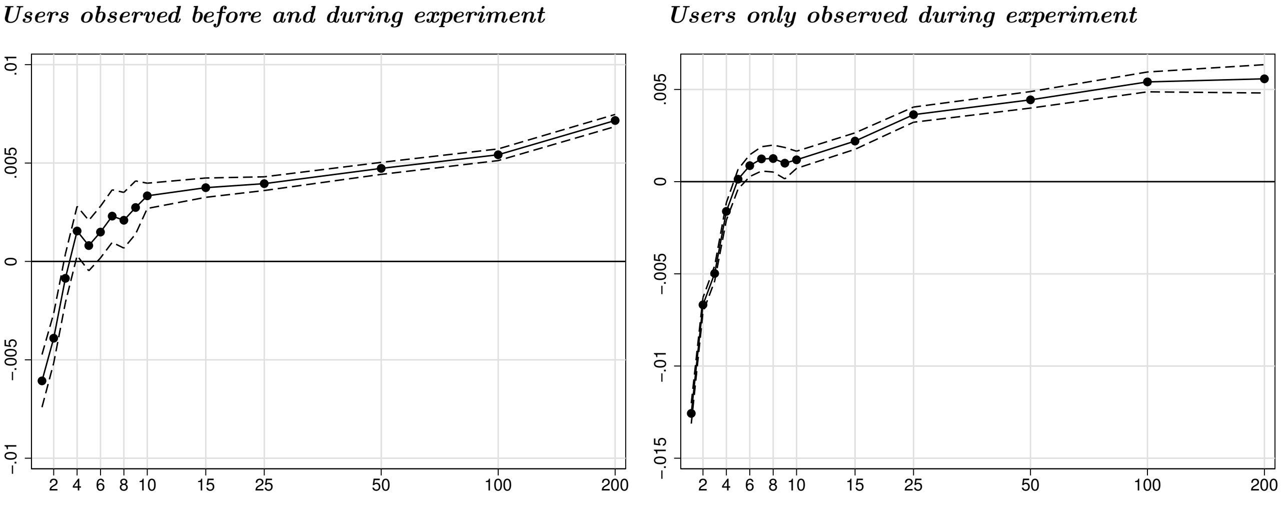
The dependent variable is the number of clicks on Slot 4. The unit of observation is user-session and the number of observations includes all individuals observed during the experimental period. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A.3: Robustness: Alternative Variation

	First Session of Morning	First Session of Afternoon	First Session of Hour	First Session of Day
	(1)	(2)	(3)	(4)
VARIABLES	Slot 4	Slot 4	Slot 4	Slot 4
Treatment	-0.0067***	-0.0065***	-0.0088***	-0.0048***
	(0.0001)	(0.0001)	(0.0000)	(0.0001)
Treatment × Prior Visits	0.0022***	0.0021***	0.0024***	0.0021***
	(0.0000)	(0.0000)	(0.0000)	(0.0001)
Individual FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Observations	67,803,071	68,911,618	95,644,773	57,704,883
R-squared	0.339	0.335	0.239	0.398

The dependent variable is the number of clicks on Slot 4. Column (1) uses data from the first session of the morning (6am to 12pm), column (2) uses data from the first session of the afternoon (12 pm to 6 pm), column (3) uses data from the first session of every hour for a user, and column (4) uses data from the first session of every day of a user. Robust standard errors in parentheses clustered at the user level. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Figure A.1: Decreasing Returns to Data



Users observed before and during experiment

Users only observed during experiment

The figure plots the coefficients δ_q along with 99% 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.

Table A.4: 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.5: 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.4. White-robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.