PNAS
www.pnas.org

Supplementary Information for

Algorithmic Amplification of Politics on Twitter

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer and Moritz Hardt

Correspondence to
Ferenc Huszár. E-mail: fhuszar@twitter.com
Luca Belli. E-mail: lbelli@twitter.com

This PDF file includes:

Supporting Information Text

1. Materials and Methods

A. The Timelines Quality Holdback Experiment. Twitter has maintained the randomized experiment described in this paper, known as the timelines quality holdback experiment, since June 2016. The experiment allocates accounts to either a control (1% of users) or treatment (4%) group completely at random.

A.1. Assignment to Treatment or Control. Accounts were randomly assigned to treatment or control either at the experiment's onset (if the account existed at the time), or at the time the account was created (if the account was created since the experiment started). In both cases inclusion in the experiment and assignment to treatment or control were determined completely randomly. The assignment is maintained over the lifespan of the account, although users in the treatment group have the liberty to temporarily switch off algorithmic recommendations (as explained below). For the purposes of this study, individuals who voluntarily turn off algorithmic recommendations continue to be identified as being in the treatment group.

A.2. Number of users in the Experiment. The experiment included 5% of all accounts globally, of which 20% are assigned to control, and 80% are assigned to the treatment group. This amounts to tens of millions of unique user ids including those of dormant accounts and bots, only a fraction of which correspond to active users who used the product during the time period we studied. In the second quarter of 2020, Twitter reported 186 million monetizable daily active users (mDAU) (1) of which about 5% (9.3 million) were included in our study.

A.3. Tweet Selection and Presentation in Control Group. Users assigned to the control group experience the Twitter home timeline much the same way it worked before the introduction of algorithmic ranking in 2016 [*]. The timeline displays Tweets authored by accounts the individual follows (referred to as in-network Tweets) or Tweets reTweeted by accounts the individual follows (ReTweets). Tweets are ordered in reverse chronological order, showing the latest Tweet first, taking into account the timestamp of either Tweet creation (for original Tweets) or reTweet time (for ReTweets). A heuristic de-duplication logic is applied to Tweets which are reTweeted by multiple accounts the individuals follow. Due to the simplicity of Tweet selection and ranking logic, this version of the timeline is often referred to as the non-algorithmic timeline.

A.4. Tweet Selection and Ranking in Treatment Group. Users in the treatment group for this experiment experience the Twitter Home Timeline the same way that the majority of Twitter users - not included in the experiment - do. Being included in the treatment bucket of this experiment has an influence on whether the same user can be selected as part of the treatment group in other randomized experiments in which related features are tested.

Treatment users who use the Twitter app on iOS or Android devices can choose between viewing the "top Tweets first" or the "latest Tweets first" in their Home timeline[†]. By default, the "top Tweets first" option is enabled and the setting reverts back to this default after a period of inactivity. On other platforms, users do not have an option to view latest Tweets first. When the Home timeline is set to display 'latest Tweets first', it works similarly to the traditional reverse chronological timeline which users in the control group experience. When 'top Tweets first' mode is selected, the Tweets are algorithmically selected, filtered and ranked. The selection and ranking of Tweets is influenced, in part, by the output of machine learning models which are trained to predict whether the user is likely to engage with the Tweet in various ways (like, reTweet, reply, etc). The machine learning methods and ranking algorithms make these predictions based on a combination of content signals such as the inferred topic of the Tweet as well as behavioural signals such as past engagement history between the user and the author of the Tweet.

In addition to Tweets and ReTweets from accounts the individual follows, the personalized timeline may also contain 'injected' content from outside the user's immediate network. A common type of such injected content are Tweets liked by someone the individual follows. Such Tweets would not appear on chronological timelines, only if someone the user follows ReTweets (rather than just likes) the Tweet in question.

A.5. Services influencing Content in both Control and Treatment. Several services and products might influence content displayed on both Control and Treatment Timelines.

Promoted Tweets: The timeline might contain Promoted Tweets, a form of advertisement, which are marked by a label "Promoted by [advertiser account]" making them clearly distinguishable from organic content. These Tweets are not usually from accounts the individual follows. These promoted Tweets are selected by algorithms which rely on machine learning algorithms and a generalized second price auction mechanism. Although the ad models might behave differently for ranking and control users, the algorithms selecting content have no explicit knowledge of whether a user is in either group.

Restricted or Removed content: content deemed to violate Twitter's Terms and Conditions, pornographic content, content displaying gore and violence, spam and fraudulent content, Tweets containing misleading or false information might be completely hidden, displayed with a warning label, or remain hidden until the user dismisses a warning message. These limitations tend to apply to individual Tweets or all Tweets from certain Twitter accounts, and apply the same way every time the Tweet.

[*] https://help.twitter.com/en/using-twitter/twitter-timeline

[†] https://blog.twitter.com/official/en_us/a/2016/never-miss-important-Tweets-from-people-you-follow.html

Blocking and Muting: Users have the ability to Mute or Block one another. Individuals who Mute another account will no longer see Tweets from the muted account, even when it is reTweeted or replied to by someone they follow. Blocking works the other way around, if an individual is blocked by an account, they can no longer see or interact with content from that account.

A.6. Machine learning models used for ranking. The ranking content on the Home timeline is influenced by the output of deep learning models, trained to predict various types of engagements with Tweets (likes, reTweets, replies, etc) [‡]. The models receive as input various signals which broadly fall into the following categories: content features (such as the inferred topic area of the Tweet, whether it contains an image), about the engaging user (such as their engagement history or users they follow), about the Tweet author (such as engagements with their past Tweets), and about the relationship between the user and the Tweet's author (such as frequency of past engagements between them). In addition to the predictions made by these machine learning models, the final ranking of Tweets is influenced by heuristics and the output of other machine learning algorithms.

The models are trained on data from users in this experiment's treatment group as well as most users outside of this experiment (the remaining 95% of accounts). Data from individuals in the control group are not part of the training dataset of these models. However, the predictions the algorithms make are routinely evaluated on data from control users, and this is used to inform decisions about the deployment of certain features.

A.7. Other forms of personalization users are exposed to. Being assigned to the control or treatment group of this experiment only affects algorithmic personalization of the Home timeline. Users receive content recommendations or are served personalized results in other product surface areas including the "Explore" tab, Search, Trends, Events, push notifications and Tweet or conversation detail pages. In addition to this, users might receive personalized content recommendations via email or push notifications, as well as 'Who to follow' account recommendations. Therefore, it is incorrect to say that the control group received no algorithmic recommendations. Assignment to control only restricts the use of personalization in the Home timeline, which is the component users spend most active minutes in.

B. Obtaining Legislators' Twitter details .

B.1. Selection criteria for countries . When studying amplification of content from national legislators, our aim was to include data relating to as many legislatures globally as possible. We identified countries to include in our analysis based on the availability of data, in particular based on the following criteria:

Availability of data on politicians' Twitter accounts: It was possible for us to identify Twitter usernames and associated party affiliations for a large number of current legislators using either official government sources, Wikidata (2), or publicly available Twitter lists curated by the corresponding political parties or other official accounts. We used the following Wikidata query to list the number of Twitter accounts associated with each legislature.

Sufficient Twitter user base in the country: The number of unique Twitter accounts in the control group which had the relevant country code associated with their account at the time of entering the experiment was at least 100,000 supporting robust statistical analysis.

Screening for the above two criteria we identified the following list of countries (in decreasing order by number of unique users in our experiment) United States, Japan, United Kingdom, France, Spain, Canada, Germany, and Turkey. Further analysis of Wikidata entries for legislators in Turkey showed that while there was almost complete data on legislators serving in the previous, 26th term, for the current, 27th term, we were only able to identify 110 usernames for a total of 600 members. Thus we did not include Turkey in our analysis. The following countries met the first, but not the second selection criterion, and were thus not included: Sweden, Denmark and the Netherlands. The following countries met the second, but not the first requirement and were thus not included: India, Brazil, Mexico, Saudi Arabia, Indonesia.

We excluded federal or international bodies such as the European Union Parliament from our analysis and focussed on national legislatures only. Members of the EU legislature Tweet in different languages and address audiences in different countries whose activity and usage of Twitter may not be homogeneous enough, complicating the analysis and interpretation of findings.

B.2. Collecting data from Wikidata. To identify members of the current legislative term in each country we followed a variation of the following process: We identified the appropriate subclass of the 'legislator' Wikidata entity (Q4175034) which describes membership in each legislature we studied. For example, the Wikidata entity Q3044918 denotes the position "member of the French National Assembly". These entities linked to the Wikidata entity of each person who have held the title via the Wikidata property P39. "position held". This allows us to identify Wikidata entities of all current and past members of the legislature.

In countries where such Wikidata entities are available, we identified the entity describing the current term of legislature. These entities are linked to the title held via the Wikidata qualifier P2937. For example the Wikidata entity describing the 15th, current, term of the French legislature is Q24939798. These entities can then be used to identify current members of the legislature.

We additionally discarded individuals whose membership of the corresponding legislature had an "end date" (P582) qualifier, as this indicates that the person no longer holds the position. In some countries, like the United States, this is the only way to find current members of the Senate as there are no Wikidata entities for legislative terms. Where available, we can retrieve

[‡] Using Deep Learning at Scale in Twitter's Timelines.

legislators' Twitter screen names which are linked to the individual's Wikidata entity via the P2002 Wikidata property. In some cases, it was also possible to retrieve numerical Twitter identifiers. Where available, numerical IDs are more reliable as users can change their screen names. We found that legislators change their twitter screennames quite often, for example, to include "mp" or "MdB" signifying their membership of Parliament or the German Bundestag as they are elected.

We also retrieved individuals' gender or sex, which is linked to a person's entity via the Wikidata property P21. We only used gender information to ensure that members of all genders are fairly represented in our sample of Twitter accounts.

Examples of SQPARQL queries we used in each country can be accessed here: Germany, Spain, France and Sweden.

B.3. Collecting data from public Twitter lists. In addition to Wikidata, we have obtained lists of Twitter accounts of legislators in the United States and Germany from publicly available Twitter lists curated by Twitter (@TwitterGov) or political parties and political organizations in Germany (US House, US Senate, Bundestagsabgeordnete, CSU MdBs, MdBs 19.WP, MdB DIE LINKE 19. WP). We manually verified that the lists we chose were of high quality, accurate and up-to-date. We removed accounts which belong to parties, local party organisations or campaign groups rather than individual legislators.

B.4. Collecting data from official government websites. To obtain a high-quality list of Twitter handles for Members of the United Kingdom Parliament we scraped the official Parliament website (UK Parliament: MPs and Lords). The contact page of each MP may contain their Twitter handle, as well as their party affiliation. For the United States we also accessed a list of representatives and senators from Ballotpedia (Ballotpedia "List of Current Members of US Congress"). We used this list to associate the accounts found on the Twitter lists to their political parties. The Ballotpedia information was also used to identify a handful of accounts that were not already included in the curated lists. For Canada, a CSV of current MPs was downloaded from the official House of Commons website (House of Commons Canada: Current Members of Parliament). The accounts were then manually annotated resulting in a very high quality dataset in terms of coverage and accuracy.

B.5. Manual data validation and annotation. While in most countries we were able to identify Twitter details of over 70% of all representatives following the automated methods described above, and in some countries this coverage was particularly high, we cannot be certain that we did not miss individuals. Our goal was to ensure that when we miss accounts, these do not result in poor representation of certain minority groups in our dataset. We have therefore focussed manual annotation effort on ensuring that accounts of legislators who belong to certain underrepresented groups, such as women, and people of colour, are included in our dataset. In most countries, we were able to retrieve gender labels from Wikidata to aid with his process.

B.6. Groupings of parties. To test various hypotheses about the types of political parties algorithms might amplify more, we make some direct comparisons between parties in each country. The first comparison we present in Fig. S1. A compares the mainstream political left with the mainstream political right. We used a number of heuristics to determine how to make this comparison in each country. In all countries except France, the parties being compared have the largest representation in their corresponding legislature. We rely on the 2019 Chapel Hill Expert Survey (3) and Wikidata annotations to determine the ideological position of each party. Typically, Wikipedia and other public sources describe one of these parties as left-wing or centre-left, and the other as right-wing or centre-right, making the comparison unambiguous. In France, the largest parliamentary group, the governing LREM is a centrist, big-tent parliamentary group, while the parties more traditionally considered as France's left-wing and right-wing (Socialists and Republicans) are currently both in opposition. In Continental European countries, the left-wing parties compared (SPD, Parti Socialiste and PSOE) are all members of the Progressive Alliance of Socialists and Democrats in the European Parliament. Likewise, the right-wing parties (CDU/CSU, les Républicains and Partido Popular) are part of the European People's Party.

In Fig. S1B we compare far-left and far-right political parties with mainstream political parties from the same country. We selected political parties where Wikipedia entries mentioned an association with far-left or far-right ideologies, or where the 2019 Chapel Hill Expert Survey (3) indicated an extreme ideology (above 9 or below 2). These were the Japanese Communist Party (Japan far-left), La France insoumise (France far-left), Rassemblement National (France far-right, represented together with associates as Non-Inscrits in the French National Assembly), VOX (Spain, far-right), Die Linke (Germany far-left) and AfD (Germany far-right). We compared each of these parties or parliamentary groups to the largest mainstream right-wing or left-wing political party in their respective countries.

In Fig. S1C we compare governing vs opposition parties. In the United States, we consider Republicans to be the governing party and Democrats (and democratic-aligned independents like Bernie Sanders) as opposition. In Japan, the government is formed by LDP and Komeito, all other parties are considered opposition. In the United Kingdom we compare the Conservative Party against the Labour Party, according to their official designation as Her Majesty's Government and Her Majesty's Most Loyal Opposition, respectively. In France we consider LREM as well as their confidence-and-supply partners MoDem and Agir as the government, and all other representatives except EDS, which Wikipedia lists as neutral, as opposition. In Spain we consider PSOE, Unidas Podemos and their supporting Basque Nationalist Party as governing, and Partido Popular, ERC and VOX as opposition. In Germany we consider CSU/CDU as governing and everyone else as opposition. In Canada we compare the Liberal Party against the Conservative Party according to their official designation as Her Majesty's Government and Her Majesty's Loyal Opposition, respectively.

B.7. Political changes. It is common for individual legislators' party or parliamentary group affiliation changes during a legislative term. This can be due to individual resignations, party mergers or the formation of new parliamentary groups or parties. For example, the French Écologie Democratie Solidarité (EDS) parliamentary group was formed in May 2020 by members of

the governing La République En Marche! (LREM) (4). In the Japanese House of Representatives, the Party of Hope (Kibō no Tō) merged with the Democratic Party to form a new party called DPFP in 2018, the majority of DPFP representatives then joined the Constitutional Democratic Party (CDP) in September 2020, while a smaller DPFP continues to exist. Where possible we validated and updated party membership data so as to best reflect major parties throughout the study period (1 April - 15 August 2020). In France, we considered EDS as separate from LREM. We repeated our analysis considering EDS representatives as members LREM, but findings do not change qualitatively. As can be seen in Fig. S1A, EDS and LREM are very similar in terms of group amplification.

C. Media bias ratings .

C.1. AllSides Media Bias Ratings. To study exposure to politically biased media sources we obtained media bias ratings for news sources from AllSides (5). While the AllSides dataset includes news sources with a global audience (such as the BBC, Guardian, Al Jazeera, etc) it focuses primarily on the U.S. media landscape, and the media bias ratings relate to how the media bias of these sources are perceived in the United States. This dataset assigns each publication or media source into one of 'Left', 'Lean Left', 'Center', 'Mixed', 'Lean Right' and 'Right'. The data we had also contained crowdsourced judgments as well as confidence ratings which were ignored. We discarded the 'Mixed' category as it included aggregator websites and media bias rating platforms (like AllSides.com itself) rather than media sites creating original content. We have further excluded sites like Yahoo News, Google News, as well as podcasts, content studios and activist groups whose original content was not clearly identifiable or attributable. To qualify for our analysis, the content from the publication source had to be clearly identifiable on the basis of URLs shared by users on the platform, we discarded publications without a clearly identifiable URL structure.

C.2. Ad Fontes Media Bias ratings. We also obtained Media Bias ratings from Ad Fontes Media (6). This dataset contained a numerical media bias rating for each news source ranging between -38.5 (most extreme left bias) and 38.5 (most extreme right bias). In line with how the data is presented on the media bias chart [§] we discretized these numerical values into 5 intervals: Left ($x < -16.5$), Skews Left ($-16.5 < x < -5.5$), Neutral ($-5.5 < x < 5.5$), Skews Right ($5.5 < x < 16.5$) and Right ($16.5 < x$). We excluded news sources such as TV channels or programs whose content was not clearly identifiable as URLs shared on Twitter. We found that, with a few exceptions, the set of publications with an Ad Fontes rating was a subset of those with an AllSides rating.

C.3. Mapping domain names and URLs. The AllSides dataset contains websites (domain names) as well as details of Twitter accounts (usernames) associated with most publications they rated. One could therefore study exposure to Tweets from the official Twitter accounts (i.e. Tweets by @foxnews or @nytimes) or exposure to Tweets containing a link to content from each publication (i.e. Tweets containing a link to a Fox News of New York Times article). We chose to focus on URLs, because publications use their official Twitter accounts in different ways, and because Tweets from these official accounts are responsible for only a small fraction of all content impressions on Twitter. Thus, we chose to use domain names as the primary identifier of each publication. We manually added missing domain names and updated or corrected outdated entries to ensure that all major news sources had an up-to-date domain name.

C.4. Regular expressions . To identify URLs that link to articles from each publication, we created regular expressions , which were matched against the text of the URL (not the text of the website). For the majority of publications these regular expressions were based on the domain name only, catching any link to any content on the corresponding website. Several news sources publish large volumes of non-politicised content such as sports, recipes, games or weather forecasts. To filter these out and focus our analysis on news and politicised content only, we created custom regular expressions, which select only articles from certain sections of each publication only.

While the naming of sections and the URL template changed from publication to publication, we aimed at including sections which most likely contained coverage of local and global news, global health, COVID-19, politics, elections, climate, science and technology. Similarly, we excluded sports, wellness, food, or weather related sections. Custom regular expressions also allowed us to distinguish between editorial and news articles from the same publisher. These often had different AllSides media bias ratings for the same newspaper group. For example, in the AllSides dataset, the online news from Fox News is rated 'Lean Right' while Fox News opinion pieces are rated 'Right'. Similarly, the New York Times news section is rated 'Lean Left' while New York Times opinion section is rated 'Left'. For publications where editorial content was rated differently, we tried to identify editorial content based on the URL string, using regular expressions. Unfortunately, this was not always possible. For example, all URLs for the Wall Street Journal were of the form 'wsj.com/articles/[a-z0-9-]+', making it impossible to identify opinion pieces based on the URL text only. In such cases, we did not distinguish between Editorial and News content, used all articles from the domain in our analysis, and assigned them to the AllSides rating of the non-editorial content (which tended to be less partisan).

We then identified Tweets with content from these publications by screening public Tweets created between 1 March and 30 June 2020, and matching any URLs these Tweets contained against the regular expressions we curated. The resulting dataset contained AllSides annotations for 100,575,284 unique Tweets pointing to 6,258,032 different articles and Ad Fontes annotations for 88,818,544 unique Tweets pointing to 5,100,381 different articles.

[§] https://www.adfontesmedia.com/interactive-media-bias-chart/

C.5. Comparison of AllSides and Ad Fontes ratings. To aid the interpretation of data, we have compared AllSides and Ad Fontes ratings for sources where we had both ratings available. Figure S3 shows the number of news sources for each pairing of Ad Fontes and AllSides rating. Figure S4 shows the breakdown in terms of number of Tweet impressions. Table S1 gives examples of specific news sources that have a certain combination of Ad Fontes and AllSides rating.

D. Measuring amplification.

D.1. Impression events. Our measures of amplification are based on counting events called "linger impression": these events are registered every time at least 50% of the area of a Tweet is visible for at least 500 ms (including while scrolling). We note that these impression events are different from what is often called a 'render impression' in the context of online media, which is registered every time the Tweet is rendered in the user's client, or every time a Tweet is fetched by the client on the user's device. For example, when the user scrolls through a large volume of Tweets in rapid succession, without stopping to allow time to read or see any of the Tweets, several render impressions would be registered, while few or no linger impressions are logged. Linger impressions are the best proxy available to us to tell if a user has been exposed to the content of a Tweet. We note that this definition of a linger impression is not the authors' choice, and since it is hard-coded into the client software, it was not possible for us to consider alternative definitions of impression.

For the purposes of this paper we considered linger impressions registered on Android, iOS, desktop web clients. The timeline's behaviour is implemented similarly in these platforms, and together they represent the overwhelming majority of user activity. We only considered linger impressions registered in the Home timeline component (the same users might encounter Tweets in other product areas such as Search, Trends, etc, but these impressions are not considered here). For each linger impression a country code is inferred from the user's IP address at the time the event is registered. When analysing content from legislators, we further restrict impressions from the relevant country (for example when considering impressions of Tweets of French legislators, we only count impressions from France). When analysing content from publishers included in the AllSides media bias ratings, we restrict impression events from the United States.

D.2. Defining amplification. Let T denote a set of Tweets. Let $U_{control}$ and $U_{treatment}$ denote the control and treatment group of users in the experiment, respectively. Note that, in our experiment, $|U_{treatment}| = 4|U_{control}|$. Let $U_{t,d}$ denote the set of users who registered a linger impression with Tweet t on day d. For a set of Tweets T, we further define $U_{T,d} = \bigcup_{t \in T} U_{t,d}$, the set of users who encountered at least one Tweet from T on day d. We define the amplification of the set of Tweets T on day d as:

$$a_d(T) = \left(\frac{|U_{T,d} \cap U_{treatment}| + 1}{4|U_{T,d} \cap U_{control}| + 4} - 1\right) \cdot 100\%$$

Often, we consider amplification within a specific country. In this case we use $U_{t,d,c}$ in place of $U_{t,d}$, where $U_{t,d,c}$ denotes the set of all users who have registered an impression event involving Tweet t while using Twitter from an IP address that we identified to be within the country c.

This gives rise to the following definition of amplification within country c on day d:

$$a_{d,c}(T) = \left(\frac{|U_{T,d,c} \cap U_{treatment}| + 1}{4|U_{T,d,c} \cap U_{control}| + 4} - 1\right) \cdot 100\%$$

For brevity, we will ignore the subscripts c and d in the discussions that follow, but note that amplification is always calculated on a daily basis, and within a specific country. When we talk about the amplification of a group G of individuals, such as members of a political party, we mean the amplification of the set of all Tweets created by this group T_G. The amplification for a group of users G is therefore:

$$a(G) = a(T_G)$$

Analogously, when referring to the amplification of an individual user i, we calculate this based on the set of Tweets, T_i, or equivalently, the group amplification for the singular set containing only i, that is $a(i) = a(\{i\})$.

When evaluating the amplification of news media, we consider the group of Tweets T which contain a URL link that we identify to originate from a particular source, or a group of news sources, and calculate the amplification of this group of Tweets.

E. Comparing amplification between groups of individuals.

Let's say we would like to compare two groups, G_1 and G_2, of users in terms of whether amplification favours group G_1 or G_2. Based on the amplification metrics we defined above we have two main ways of defining "equal" amplification.

E.1. Equal group amplification. Since we have a way to measure the amplification of a group G of individuals, expressed in terms of the amplification for the set of Tweets T_G, we can simply say the two groups are equally amplified if:

$$\bar{a}(G_1) = \bar{a}(G2) \quad [1]$$

where $\bar{a}$ is the average daily amplification. In our paper we calculate amplification on a daily basis over a 90 day period. Note that this comparison is not robust to outliers within groups G_1 and G_2. If the groups contain very active individuals whose Tweets are seen more than others, their data would dominate the measure of amplification for the group. To increase the robustness to outliers we propose the following criterion: Let $\tilde{G}_1$ and $\tilde{G}_2$ be bootstrap resampled variants of G_1 and G_2

respectively. That is, $\tilde{G}_1$ is a random subset of G_1 formed by sampling uniformly from G_1, $|G_1|$ times. We can say that group G_1 and G_2 are amplified equally if the following condition holds:

$$P[\bar{a}(\tilde{G}_1) > \bar{a}(\tilde{G}_2)] = P[\bar{a}(\tilde{G}_1) < \bar{a}(\tilde{G}_2)] \quad [2]$$

It is possible to estimate $P[\bar{a}(\tilde{G}_1) > \bar{a}(\tilde{G}_2)]$ from data, and to determine whether it is significantly different from $P[\bar{a}(\tilde{G}_1) < \bar{a}(\tilde{G}_2)]$ by a permutation test.

E.2. Individual amplification parity. Another option is to calculate the amplification $\bar{a}_i$ for any individual i in group G_1 and G_2 and compare the distribution of amplification values between the two groups. A strong criterion would require statistical independence between amplification and group membership, requiring the distribution of individual amplification values to completely agree between the groups. This is a very strong requirement and it is difficult to reliably establish independence when groups are small. Instead, we use a weaker notion of equivalence as follows. If I_1 is a random individual sampled uniformly from G_1 and I_2 is a random individual sampled uniformly from G_2 we say that the amplification of individuals in G_1 and G_2 is essentially equal if:

$$P[\bar{a}(I_1) > \bar{a}(I_2)] = P[\bar{a}(I_1) < \bar{a}(I_2)] \quad [3]$$

Similarly to the bootstrap-based criterion for equal group amplification, we can estimate $P[\bar{a}(I_1) > \bar{a}(I_2)]$ from data, and determine whether it is significantly different from $P[\bar{a}(I_1) < \bar{a}(I_2)]$ using a permutation test.

E.3. Relationship between equal group amplification and individual amplification parity. In this section we discuss the relationship between comparing groups based on group amplification or the distribution of individual amplification values. It is easy to see that if amplification $a(G)$ of a group G were a linear function of the amplification of individuals $i \in G$, that is, when $a(G) = c_1 \cdot \sum_{i \in G} a(i) + c_2$, then individual amplification parity (Eq. (3)) implies equal group amplification (Eq. (2)).

However, our definition of amplification does not satisfy this requirement. To see why, consider the function $f(G) = |U_{T_G}|$, where T_G is the set of Tweets authored by members of the group G and U_{T_G} is the set of users who registered an impression event with at least one Tweet in T_G. The function f is a submodular set function exhibiting a diminishing return property $f(G \cup H) \leq f(G) + f(H)$. Equality would hold if Tweets from groups G and H reach completely non-overlapping audiences. For most of the groups we consider, such as groups of politicians from a political party, this is highly unlikely. We define amplification as the ratio between two such submodular set functions. As a consequence, equivalence of amplification at an individual level does not imply equivalence at the group level, and vice versa.

E.4. Identifying outliers via leave-one-out analysis. To find the most significant outliers in each group in terms of amplification, we performed a leave-one-out or jackknife analysis. We calculated group amplification for each group with each one of the members left out. We then selected positive and negative outliers on the basis of how much leaving the member out of the group has changed the aggregate amplification of the rest of the groups. For a member of the group to be identified as an outlier by this process, it has to have a significant audience (so it has a substantial contribution to the overall group amplification) and it has to behave substantially differently than other members of the group.

2. An overview of the research process and Twitter's review process for publications

This team of authors carried out the research reported here with considerable autonomy and independence throughout the research process. Our findings are shared without cherry picking. In this section we give a detailed account of our research process and steps taken to minimize the influence of corporate interests.

A. Hypothesis selection. Shortly after the idea of this research was conceived of (in November 2019), the authors have pre-registered a number of hypotheses which could be tested using the massive-scale experiment reported on in this paper. This pre-registration was informal, inasmuch as the pre-registration document was not made public and was not required or reviewed as part of the internal processes at Twitter. These hypotheses centred around algorithmic amplification of abusive content/hate speech, political polarization, and misinformation. No one except the authors of this paper were involved in selecting hypotheses or designing research methods.

Initially the researchers worked in parallel on testing hypotheses related to abusive Tweets and political content. In June 2020, the authors decided to prioritise research on the questions related to political content, because of the higher quality third party data that was available. There were no other considerations that influenced hypothesis selection.

B. Project prioritisation. The research project was submitted to be prioritised as a research project within the Cortex Applied Research team in December 2019. This prioritization process is used to allocate headcount to research projects. This was the first time the aims and scope of this research project was disclosed widely within the company. The project was judged to be high-impact for the company, and the resulting decision was to allocate headcount, allowing the coauthors to spend more time on the project.

C. External input to the research project. Consultation with parties outside of the immediate research team (authors of this paper) changed the direction of research in two specific ways:

- **More global scope:** Initial results focused on the United States and United Kingdom. The public policy team pointed out the bias in choosing to focus on these two countries while Twitter is used globally. Following up on this observation we extended the analysis to further countries, and we believe we included all countries where it was technically possible. An effort was made to reduce subjectivity in choosing which countries to include in our analysis: we have extended the analysis to all countries where we could reliably identify a large number of legislators, and where we had enough data from users. The resulting selection criteria are deteiled in SI Section 1.B.1.
- **Using multiple media bias rating datasets:** Initially, the team had access to only one media bias rating dataset (AllSides). Multiple teams raised concerns about the reliability of this third party source, and about making our findings dependent on the validity of a single underlying dataset. We have therefore obtained a license for a second media bias dataset (Ad Fontes Media). We present both sets of results in our paper with no normative judgment on the validity of either of these underlying sources of data.

D. Internal review process. Before a research paper is submitted to peer review (or is published in other form), an internal review is conducted where representatives of the following teams have to give approval:

- **The public policy team** reviews the paper's likely impact on public policy.
- **The investor relations team** reviews the paper to ensure that it complies with regulations regarding the release of material non-public information.
- **The intellectual property legal team** reviews if the paper discloses any non-public information relating to company IP, and whether any third-party data or intellectual property is used under correct licensing terms.
- **The communications team** reviews the paper to ensure the manuscript is consistent with broader Twitter terminology, and to be able to plan additional communication such as blog posts and Tweets from company accounts. Suggesting changes to the interpretation or presentation of findings in the paper would be beyond the scope of this review.
- **Two technical reviewers**, usually employees of the company and nominated by the authors, comment on the paper's technical contributions, scientific rigour and the quality of presentation.

For this manuscript the internal review started on 27 Aug 2020. None of the reviewers suggested any changes to the paper or made any substantial comments beyond commenting positively on the work or stating their approval. After necessary approvals were given and a full Data Protection Impact Assessment was completed, the manuscript was first shared by a journal editor in a pre-submission enquiry on 25 Sep 2020. While it is possible through this review process for Twitter to refuse a request to publish a paper on various grounds, to our knowledge this has never happened at Twitter.

E. Privacy and Data Protection Reviews. As the project involved using personal data, several privacy and data protection (PDP) reviews were conducted at various stages to ensure compliance with Twitter's policies regarding data retention and protection of user privacy. These reviews did not alter the research decisions the team made.

F. Ethical Approval. The control group assessed was not created for the purpose of research but rather for the business purpose of improving the algorithm and providing a baseline by which it could be compared to monitor the ongoing performance of the algorithm. As such, this work was reviewed by Twitter's legal and privacy teams as part of its ordinary business operations (and not an IRB). As part of this review, a data protection impact assessment was conducted, and it was determined that additional notice and consent mechanisms were not required.

3. Data availability

To allow reproduction and limited extension of our findings we make the following data available upon request.

A. Terms of data access. Data will be made available upon request, by emailing the corresponding author. Prior to accessing the data, researchers will be required to sign up for a Twitter Developer Account at developer.twitter.com. Access to the Data Set is governed by the Twitter Developer Agreement and Policy available online and prior to accessing the data set all researchers must agree to this agreement. Further sharing of data is not permitted.

B. Details of the dataset. In this section we give details of data we will make available upon request which allows the reproduction of main findings presented in Figures 1A and 1B, Figure 2, supplementary Figure S1. In addition to the comma separated data files whose details are given below, a python jupyter notebook will be provided to reproduce our specific plots from the data.

`parties_aggregate_amplification_bootstrap.csv` allows the reproduction of Figures 1A, 1B and S1, contains amplification values for each political party in one of the 7 countries studied. To increase the robustness of our findings we used bootstrap: each 'bootstrap fold' in this dataset relates to different random subset of politicians from each party, with amplification calculated over a random sample of days from within the study period. This file has 7,074 rows and the following columns:

- `grouping_id`: encodes a legislature and particular way of grouping legislators where ambiguous.
- `group_label`: identifies a party or political group.
- `bootstrap_fold_id`: identifies the bootstrap fold
- `amplification_ratio`: numerical amplification value expressed as a percentage, as reported in Figures 1 and 2. 0% indicates no relative amplification compared to control condition.

`media_bias_categories_daily_audience_bootstrap.csv` allows reproduction of Figure 2, contains the aggregate daily US audience of Ad Fontes or AllSides media bias categories between 15 April and 15 August 2020 within the two experimental conditions. This file has 185,501 rows the following columns:

- `rating_source`: either adfontes or allsides
- `media_bias_category`: identifies media bias category, such as 'Lean Left'
- `date`
- `bootstrap_fold_id`
- `experimental_condition`: either ranking or control
- `audience`: unique number of users within experiment condition who encountered at least one Tweet with a link from this category

`media_sources_daily_audience.csv`: contains more fine grained data on individual U. S. media sources, not aggregated by media bias ratings but instead listed for each individual publication. This file can be used to reproduce the outliers in Figure 2 of the main paper, and can be used for further analysis and extensions. To mitigate privacy risks associated with the readership of very small publications, we only include the top 150 media outlets we studied, these have large enough audiences to mask any individuals: the smallest daily audience measurement in this CSV is 102. This file has 30,005 rows and the following columns.

- `publication_title` e. g. AP or Al Jazeera
- `date`
- `experiment_condition` either ranking or control, identifies whether the audience is within the treatment or control group
- `audience` number of unique users in each experimental condition who have encountered at least one link from this publication/media source in the given day

`legislator_screennames.csv` and `legislator_screennames_grouped.csv` (a processed version of the same) contain the list of Twitter accounts of legislators in 7 countries with their associated party affiliations at the time of the analysis. This dataset is derived from publicly available sources and has been improved via manual data curation (see SI 1.B). While this is not needed to reproduce our results, sharing this data could be useful in validating the quality of the data underlying our analysis and to check details of each legislator's assumed party membership. This file has the following columns:

- `country_code`
- `chamber` e.g. US House vs Senate
- `group_id` name of political party or group
- `screenname`

`media_sources_regex.csv` contains regular expressions we used to extract Tweets containing a link to one of the US media sources we analyzed (See SI 1.C.4). This is not needed to reproduce any figures, but allows additional scrutiny of our methods. This file has the following columns:

- `media_source_name`
- `media_source_domain` the domain name associated with the online news source
- `is_opinion` binary value indicating whether this row relates to editorial/opinion content or regular news reporting. Used to distinguish between e. g. NY Times News and NY Times opinion as separate media sources.
- `regex` regular expression matched against a url to determine if the URL points to an article from the media source in question

C. Third party media bias data. Licensing terms do not permit us to share raw data from Ad Fontes Media or AllSides which we used in our analysis. Researchers who wish to reproduce the mapping of media sources to media bias categories are encouraged to obtain a license and the data from the AllSides and Ad Fontes Media websites.

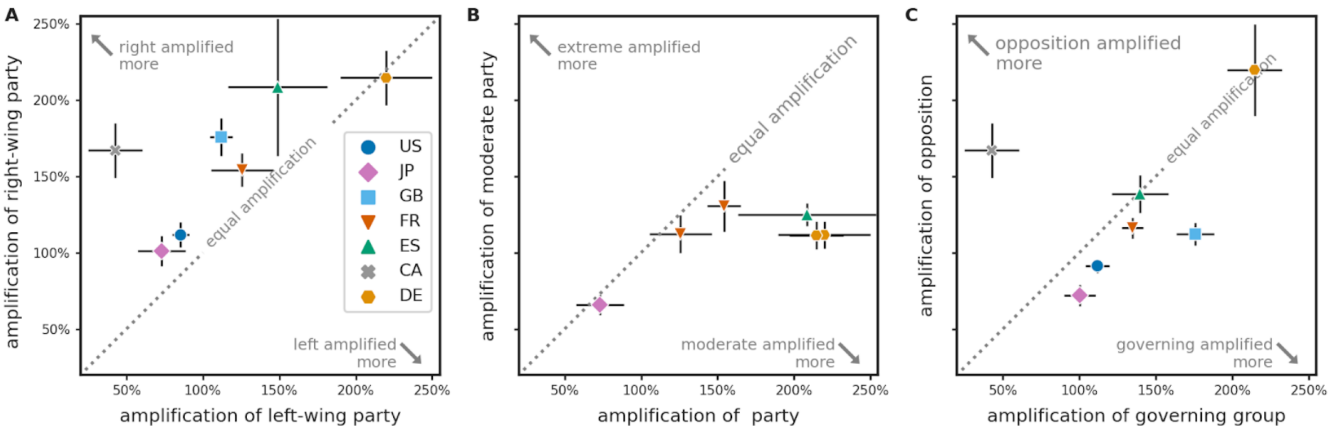

Fig. S1. Evaluating various hypotheses about algorithmic amplification of political parties. **A** (This panel is identical to Fig. 1B) Comparing the amplification of the largest mainstream left- and right-wing parties in each country: Democrats vs Republicans in the U.S., CDP vs LDP in Japan, Labour vs Conservatives in the U.K., Socialists vs Republicans in France, PSOE vs Popular in Spain, Liberals vs Conservatives in Canada and SPD vs CDU/CSU in Germany. **B** Comparing extreme far-left or far-right parties against relevant mainstream parties from the same country: CDP vs JCP (left) in Japan; LFI vs Socialists (left) and RN vs Republicans (right) in France; VOX vs Popular (right) in Spain; Die Linke vs SPD (left) and AfD vs CDU/CSU (right) in Germany. **C** Comparing the governing parties against the opposition parties in each country. In the United States Republicans were considered governing. Error bars in all panels show standard error estimated from bootstrap.

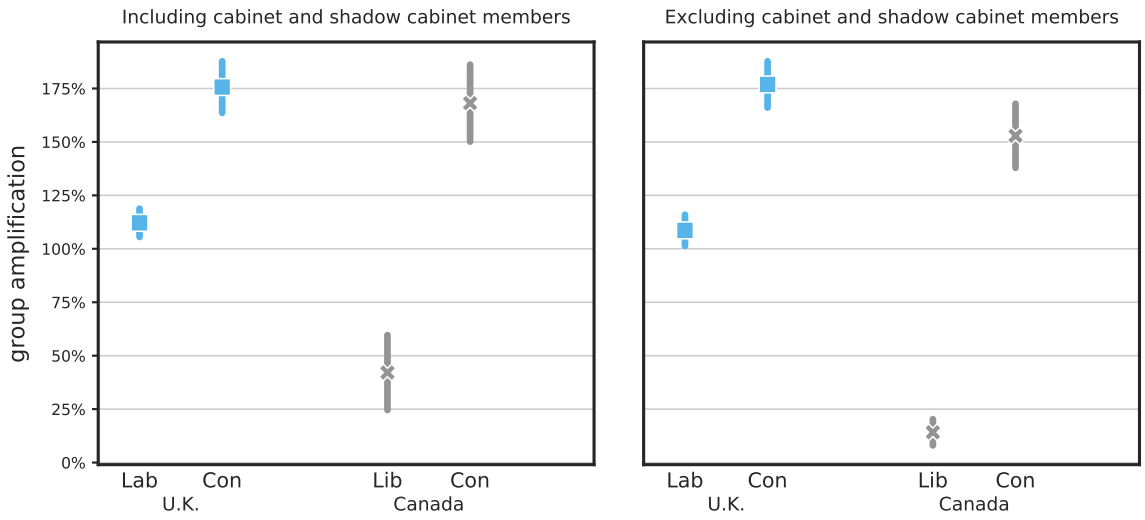

Fig. S2. Ablation study reproducing findings from Figure 1 for the U. K. and Canada with members of the cabinet and shadow cabinet removed from the analysis. *Left panel:* In these countries the Prime Minister and key members of the government are also elected Members of Parliament, they are thus included when calculating group amplification of each party in Figure 1A. *Right panel:* We rerun the analysis with members of the Cabinet and members of the Shadow Cabinet (high ranking opposition politicians) left out. Inthe U.K. we excluded Prime Minister Boris Johnson, and holders of Great Offices of the State Matt Hancock, Rishi Sunak, Priti Patel and Dominic Raab. Likewise, we excluded Shadow Prime Minister Keir Starmer and Shadow Cabinet Members Angela Rayner, Lisa Nandy, Anneliese Dodds and Jonathan Ashworth. In Canada we excluded Prime Minister Justin Trudeau and cabinet members Chrystia Freeland, Harjit Sajja, Catherine McKenna and Ahmed Hussen and opposition politicians Andrew Scheer, Pierre Poilievre, Erin O'Toole and Candice Bergen. Our findings are qualitatively similar irrespective of whether cabinet members are included or not.

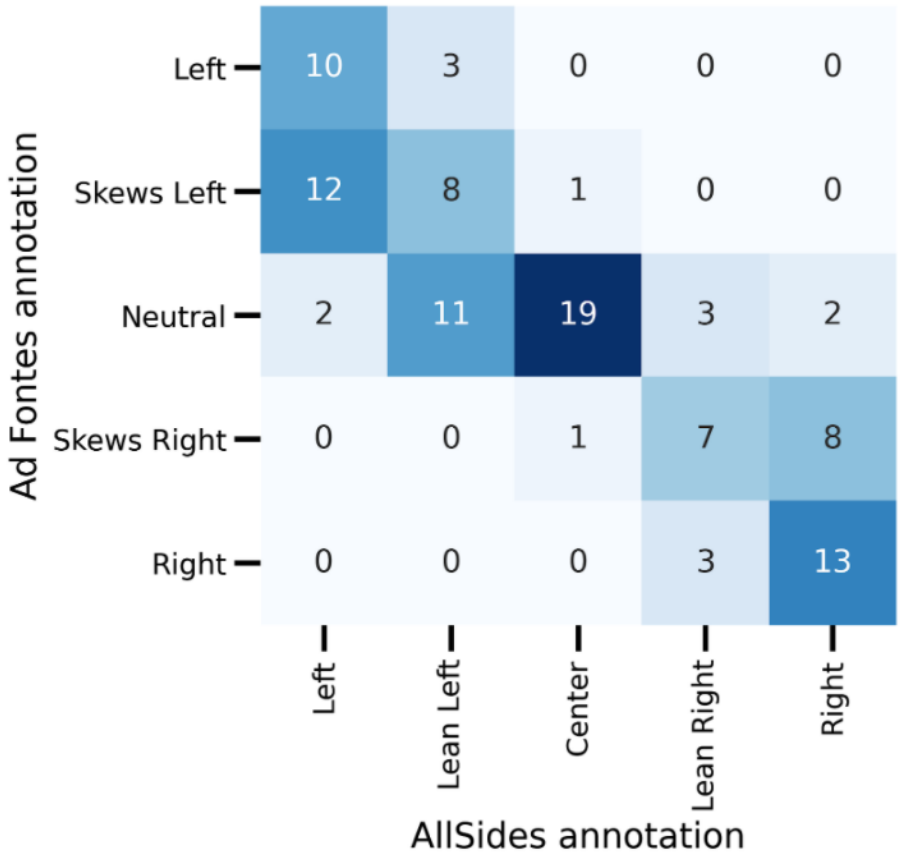

Fig. S3. Number of news sources for a particular combination of AllSides and Ad Fontes rating, for publications where both ratings were available. The comparison reveals a tendency for Ad Fontes to rate publications more neutrally compared to AllSides, especially on the Left end of the spectrum. For example, the majority of sources rated as 'Left' by AllSides is rated as 'Skews Left' by Ad Fontes. Similarly, the majority of sources rated 'Lean Left' by AllSides is rated 'Neutral' by Ad Fontes.

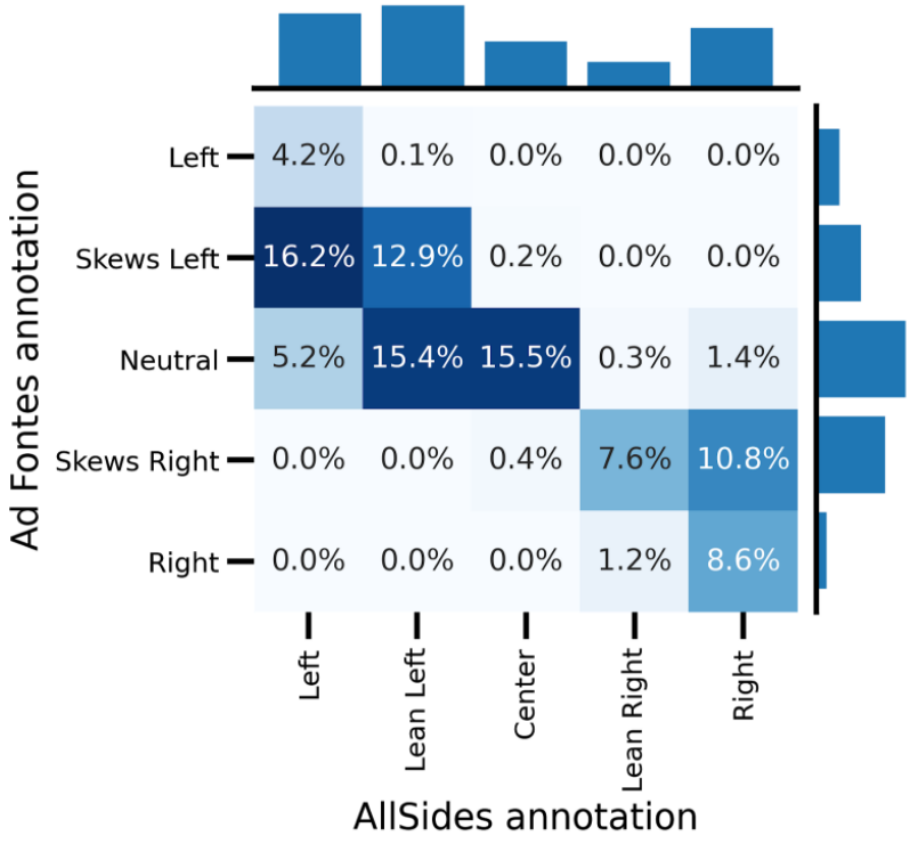

Fig. S4. Distribution of AllSides and Ad Fontes media bias ratings among Tweet impressions in the control group. The annotated heatmap shows the percentage of all Tweet impressions with a particular combination of AllSides and Ad Fontes rating. The most frequently occurring combination is AllSides Left, Ad Fontes Skews Left, accounting for 16.2% of Tweet impressions. Histograms show the marginal distribution of AllSides and Ad Fontes annotations. The most frequent Ad Fontes category is Neutral, while the most frequent AllSides category is 'Lean Left'.

Table S1. Most popular news sources for each combination of AllSides and Ad Fontes media bias rating by number of Tweet impressions in the control group.

Ad Fontes	AllSides	Top media sources by number of impressions
Left	Left	Daily Beast, Slate, Intercept
Left	Lean Left	Truthout, FAIR.org, The Advocate
Skews Left	Left	CNN (opinion), BuzzFeed, Vox, Raw Story
Skews Left	Lean Left	CNN (news), Politico, NBC
Skews Left	Center	The Week
Neutral	Left	New York Times (opinion), NY Daily News
Neutral	Lean Left	New York Times (news), Washington Post, LA Times
Neutral	Center	Bloomberg, Business Insider, Associated Press
Neutral	Lean Right	Reason, Marketwatch, Fiscal Times
Neutral	Right	New York Post, Daily Mail
Skews Right	Center	Real Clear Politics
Skews Right	Lean Right	Fox News (news), Washington Examiner, Washington Times
Skews Right	Right	Fox News (opinion), National Review, The Epoch Times
Right	Lean Right	One America News Network, PJ Media, Judicial Watch
Right	Right	Breitbart, Daily Caller, The Gateway Pundit

References

1. Twitter, Q2 2020 letter to shareholders. published by Twitter Investor Relations, Online; retrieved August-2020 (2020).
2. D Vrandečić, M Krötzsch, Wikidata: a free collaborative knowledgebase. *Commun. ACM* **57**, 78–85 (2014).
3. R Bakker, et al., 2019 Chapel Hill Expert Survey, version 2019.1., available on chesdata.eu (2020).
4. C Cornudet, I Ficek, "Le groupe En Marche en passe de se fracturer à l'Assemblée" (The En Marche group on the brink of fracturing at the Assembly). (2020) [Online; posted 08-May-2020].
5. AllSides, How AllSides Rates Media Bias. (2020) [Online; retrieved August-2020].
6. V Otoro, How Ad Fontes Ranks News Sources - Methodology Summary. (2020) [Online; retrieved August-2020].