scientific reports

Check for updates

OPEN

Social bots spoil activist sentiment without eroding engagement

Linda Li[1,2,7]✉, Orsolya Vásárhelyi[3,4,5,7] & Balázs Vedres[5,6]

Social bots are highly active on social media platforms, significantly affecting online discourse. We analyzed the dynamic nature of bot engagement related to Extinction Rebellion climate change protests in 2019. We found bots to impact human behavior more than the other way around during active discussions. To assess the causal impact of bot encounters, we compared communication histories of those who interacted with bots with matched users who did not. There is a consistent negative impact of bot encounters on subsequent sentiment. The impact on sentiment is conditional on the user's original support level, with a more negative impact on those with a favourable or neutral stance towards climate activism. Political 'astroturfing' bots induce an increase in human communications, while encounters with other bots result in a decrease. Bot encounters do not change activists' engagement levels in climate activism. Despite the seemingly minor impact of individual bot encounters, the cumulative effect is profound due to the large volume of bot communication. Our findings underscore the importance of monitoring the influence of social bots, as with new technological advancements distinguishing between bots and humans becomes ever more challenging.

Keywords Social bots, Human–bot interaction, Information cascades, Political communication, Protests

Social media has become the primary channel to engage in political discussions, and to participate in collective action over the past decade[1–3]. However, on social media there are automated agents as well, among them "social bots", that are increasingly active in our social and political publics[4], creating content and interacting with humans with ever increasing sophistication[5]. This hybrid ecosystem where algorithmic agents and humans co-exist can fundamentally alter the nature of democracy, political accountability, transparency, and civic participation[6]. Automated accounts can propagate a large volume of messages at minimal expense, and can engage with users with a fast reaction speed[7]. As a consequence, the algorithmic share of social media communications is now on par with human participation: Automated users were estimated to be responsible for generating 10–40% of tweets in recent political events, such as the 2016 US presidential election, the Brexit referendum, the yellow vests movement, the Catalan referendum, or the 2019 United Nations Climate Change Conference[8–10].

Social bots—especially those intending to mimic human behavior—can disrupt online political discussions[11], and significantly influence political debates and activism[11–13]. Such bots are frequently designed to pass as human accounts, and occasionally also mimic known political figures and government accounts to gain the attention and trust of human users[14]. Bots are often highly active during the flare-up of discussions around new political events[15], and disseminate targeted messages ranging from fabricated news to contentious, divisive, and negative content[16–18], blending legitimate messages and misinformation. Furthermore, bots are also often deployed in orchestrated efforts to generate the facade of a seemingly vibrant discussion conforming with hidden agendas[19,20] by retweeting each other (known as "astroturfing"), targeting susceptible users[8].

Although social bot presence has been studied before at the macro scale, less is known about the micro-level impact of human-bot encounters on subsequent human activism. Research on bot-human interaction found that bots can often hijack the topics and overall tone of human discussions[10,16], and they often increase the visibility of extreme views[8,16,21], influence sentiment around topics[22], and intervene in human communication flows[23,24]. Small-scale simulations and experiments indicate that bots can alter expressed human values[24] and behaviors[25], particularly driving users towards more extreme viewpoints[26], and potentially impact human's level of online activity[27] as well as offline political participation[28]. However, there is little empirical evidence regarding the capacity of bots to modify human behavior in real-life political communication[17,19,23,29].

[1]Department of Methodology, London School of Economics and Political Science, Columbia House, Aldwych, London, UK. [2]Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, UK. [3]Center for Collective Learning, Corvinus Institute for Advanced Studies, Corvinus University, Budapest, Hungary. [4]Institute of Data Analytics and Information Systems, Corvinus University, Budapest, Hungary. [5]Democracy Institute, Central European University, Budapest, Hungary. [6]Department of Network and Data Science, Central European University, Vienna, Austria. [7]These authors contributed equally: Linda Li and Orsolya Vásárhelyi. ✉email: l.li88@lse.ac.uk

Today, online activism constitutes an essential part of democracy. Online activism can be defined as the strategic use of digital communication technologies by individuals or groups to engage in political and social change efforts. These efforts include social media campaigns, online petitions, digital protests, and the diffusion of messages through digital platforms to mobilize individuals or communities around specific causes, and advocate for policy changes[30]. However, most research on social bots focuses on automated interventions in institutionalized political processes such as elections[21,22,31,32], while few studies examined the role of bots within activism related to online protests[10,23,33]. The dynamics of social media use in online activism differs markedly from communication around other political events[34–37]. Since social bots are designed to be more responsive than humans[38], it is common to observe an increase in bot activity during the peak of heated online debates, followed by a decrease in their presence[8]. While the bursty nature of social media communication of protests is well known[39,40], the impact of social bots on human activity during and after bursty periods has not been investigated.

How does interacting with social bots impact human behavior in online activism? We address this question with data on X (Twitter) discourse on climate-change-related social movements. We analyze the dynamics of bots and humans engaging with each other, and we also compare the difference in impact of direct communication with bots to activists who did not directly interacted with bots. We focus on events of direct communication with bots—when humans engage with bot messages in writing (replying, mentioning or commenting bot tweets)—and not merely events of seeing bot communications, as bot messaging today is highly common, and seeing one automated message should not have much impact. Our analysis focused on protest-related discourse during a series of protest events that erupted from November to December 2019. We decided to concentrate on online activism related to climate change as our case study, as algorithmic threats to engagement in climate change activism can have profound consequences on societal agreement on the public good in a critical issue[41]. We analyzed communication around the Extinction Rebellion (XR), as the highest profile activist group online.

The topic of climate change has been shown to attract highly engaged, active, and committed participants[42], while there is also substantial bot activity[43,44]. This enables us to analyze the impact of human-bot interactions on humans during information cascades[19], and measure the effect of bot-human interactions on tweeting activity[27] and sentiment[26]. Our findings contribute to the growing body of research on machine behaviour[6], particularly in terms of understanding how rapidly developing hybrid human-machine systems could potentially modify human opinion over an extended period.

Early work related to human–bot interactions on social media found that user sociability and network size predict who will be interacted by a bot[15,45]. Our results show that bots have become so widespread that it is unlikely that an active user of X will *not* meet a bot online. Our research extends previous work by focusing on quantifying the impact of direct human-bot communication. We found that bot type matters for the impact on users' tweeting activity, and the initial level of support a user has toward the climate change movement determines how a bot encounter impacts their sentiment on climate change. Our results have important policy implications for increasing platform transparency in how they handle automated profiles.

Results

Proportions of bot and human communication

We found 48% of all accounts to fall into the bot category within our sample from Twitter (44,121 of the total 93,499 users). We identified automation through a combined approach[46,47] that integrates the results of commonly used bot identification methods, Botometer[48] and our self-trained machine learning-based algorithms. Botometer is a widely adopted open source tool to identify bots on Twitter[49]. Our self-trained models used various data sets tailored to identify users with automated behavior that attempts to mimic humans on social networks, especially in the context of political behavior[48]. Bots reported in the main text are the combined results of these two models, with a fixed threshold for the Botometer ($CAP >= 0.65$). Since the concept of "bot" encompasses varying degrees of automation, using one fixed bot classification threshold is always a simplification. Therefore, we repeated all analyzes reported in the main text at various bot thresholds, and we report these results in the SI. (See 5, 5.2 and SI Bot Identification for more details. Supplementary Information (SI) Table S3 shows details of the training set and data used for bot identification, while Figure S1–S4 shows metrics used for fine-tuning bot detection models).

Figure 1 shows the information flow between bots and human users. 81% of tweets were replies or retweets in our database, and 51% of these retweets originated from bots. In general, bot activity is mainly the posting of original messages or the retweeting of each other (35% of all retweets and 71% of bot retweets were retweeting other bots), and only a small portion of bot retweets were retweets originating from humans (29%). At the same time, humans spread roughly the same number of messages produced by bots (54%) and humans (46%). If the bot threshold is increased to $CAP >= 0.75$, still 48% of the retweets originated from bots, and 45% of the human retweets originated from bots (see SI Information Flow and SI Tables S6 to S8 for information flows).

Temporal co-dependence of bot and human tweeting

We found that the quantity, intensity, direction and the sentiment of the information flow between humans and bots are highly topic-dependent. We identified seven topics related to XR protests associated with news events, political campaigns, or outbursts of sentiments (for example, climate change denial) by applying bi-term topic models[50]. (See 5, 5.3, Topic Modeling for more details and SI Table S5 for a brief summary of the topics, including their content, top hashtags, and sample tweets). Out of these seven topics, four were highly bursty: Topic-related activity increased sharply periodically and then decreased suddenly[51], resulting in information cascades. Information cascades occur when users follow the behavior of other individuals on social networks (e.g. retweeting the same message)[52]. (See Table S9 in SI Burstiness Scores).

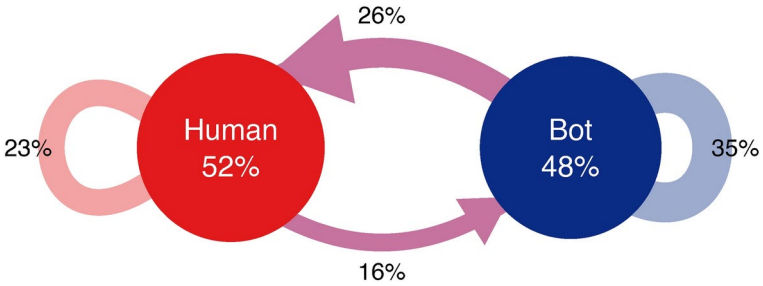

Fig. 1. Information flow of humans and bots. The directions and amount of tweets retweeted between bots and humans.

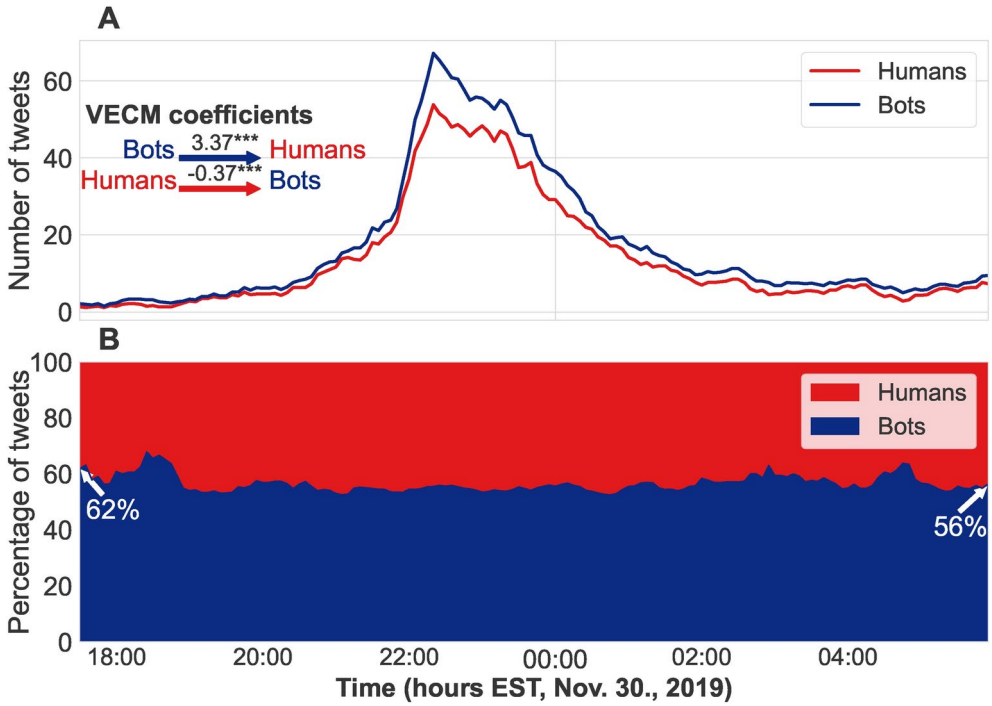

Fig. 2. Panel (**A**): Number of tweets posted by bots (blue line) and humans (red line) in a cascade mutually-driven by bots and humans within the "Anti-XR protest" topic. Panel (**B**): Ratio of bot and human generated tweets throughout the time period of the cascade. The number of tweets is the rolling average aggregated on 5-minute intervals. Bots are classified as users who has a overall bot probability no less than 0.65.

Figure 2 shows the temporal distribution of human and bot tweets aggregated at a 5-minute interval within an illustrative bursty period of climate change discussions on Twitter. Although human users (red line) generated a higher volume of tweets during the cascade's peak, the tweeting frequency trend is influenced by bot. Table 1 shows the results of Vector Error Correction Models (VECMs). VECM is a statistical model used to analyze and estimate whether one time series could be used to predict another after introducing a time lag and potential confounds[53].

We found that in 3 out of 4 identified cascades, the number of bot tweets during bursty periods could be used to predict human activity, and the sentiment of tweets by humans could predict the sentiment of bots' tweets. Bots impacted human tweeting activity significantly stronger in the case of "Disruptive engagement" (Wald = 10.06, $p = 0.0015$) and "Anti-XR protests" (Wald = 45.36, $p = 0.000$), but tweeting activity related to "Politicized Activism" was mutually driven by both bots and humans. (Wald = 0.76, $p = 0.39$). Humans impacted bots' sentiment significantly more negatively in the case of "Anti-XR protests" (Wald = 3.66, $p = 0.048$). In the case of "Politicized Activism", sentiment was also mutually driven by both bots and humans, with no significant difference in coefficients strength (Wald = 1.65, $p = 0.198$). The effect remains consistent across bot CAP thresholds ranging from 0.50 to 0.75 for 3 of the 4 cascades identified. (See more details on Burstiness in Data and Methods, Identifying cascades, SI Cascades, SI Tables S9–S20 for Results across different CAP thresholds, different time lags, for full time period of our analysis instead of cascades, and for Topic Burstiness Scores).

Topic	Amount bots⟶humans	Amount humans⟶bots	Wald test	Sentiment bots⟶humans	Sentiment humans⟶bots	Wald test
"Football game protests"	−0.14	0.06	1.72	−0.01	0.13	2.05
"Disruptive engagement"	57.89***	0.01	10.06**	−0.04	0.34**	2.70
"Anti-XR protests"	3.37***	−0.38***	45.36***	−0.08	−0.65***	3.66*
"Politicized activism"	0.54**	0.17***	0.76	5.09***	4.12***	1.65

Table 1. Vector error correction models (VECM) testing the relationship between the amount and sentiment of bot and human communication with 20 minutes time lags. $^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. VECM-s were used to determine whether bot activity predicts human activity (Bots to Humans) or human activity predicts bot's (Humans to Bots) in the identified four cascades. Wald tests were performed to test whether coefficients of Bots to Humans and Humans to Bots predictions are significantly different. Modelling were based on 30-minute time lags.

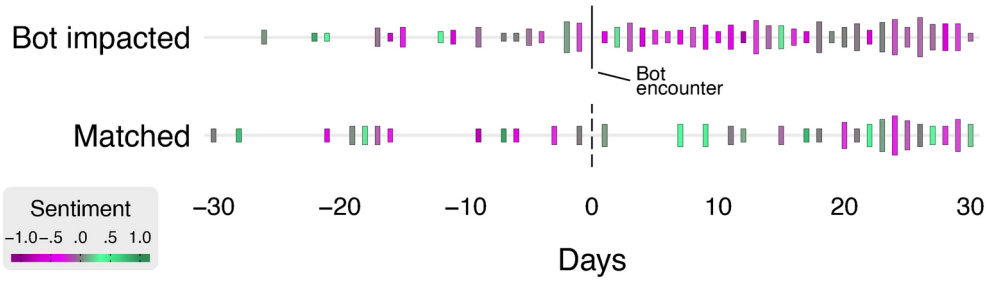

Fig. 3. Two examples of Twitter timelines. Top timeline shows daily tweet counts and daily mean sentiment for a human account that directly interacted with a bot on day 0; bottom timeline shows a matched account without a direct bot interaction. Bar heights are proportional to tweet counts, color indicates sentiment from -1 to 1.

Predicting the impact of direct bot interactions

To test whether direct communication with bots (replying, mentioning or commenting bot tweets) has any influence on human users—beyond their above shown immediate collective effect in bursty periods—we developed two models focusing on (1) Amount and (2) Sentiment. Specifically, we analyzed how human communication evolves over a span of 30 days after the first direct interactions with a bot in the discussion related to climate change on Twitter. The first model captures the inclination to "speak out" quantified by the average number of tweets posted related to the XR protest. The second model predicts the change in sentiment about the climate change protest (measured on a scale ranging from -1 to 1, with -1 representing the most negative sentiment and 1 the most extreme sentiment).

To quantify the impact of direct bot interaction on human communication change, we used a three-step process. This includes: (1) we sampled $N = 303$ users who directly interacted with bot accounts (bot-exposed)—replied or commented to a tweet/comment originated from a bot; (2) we collected a matched sample of $N = 184$ users, who were active in the XR related protest discussion on Twitter but had no direct interaction with bots. Matching users were selected on the basis of the similarity score calculated pairwise to our original sample considering publicly available metrics on human users' profiles. (See Data and Methods, Matching sample for more details.) (3) Figure 3 shows two example timelines, one human user who directly interacted with a bot (the interaction is shown at time 0), and a matched user that did not encounter a bot.

(3) Finally, we applied Difference-in-Difference (DiD)[54] regression models to quantify the casual effect of bot interaction on outcomes by comparing a set of humans who directly interacted with bots (human replying or commenting bot tweets) with those who did not. Our observation units are the daily activity of human users relative to the time of interaction with bots - 30 days *before* and 30 days *after*. This setting allows us to quantify the prolonged impact of bot interaction on the frequency of tweeting and sentiment of tweets compared to users who did not meet a bot.

Botometer provides prediction values for belonging to seven subbot categories ranging from 0 to 5. We classify a bot into a subcategory if its bot-specific probability is greater than 2.5. Our biggest group is 'miscellaneous other bots' which are bots that are similar to various type of manually annotated bots (23%), followed by manually labeled political bots, so-called 'astroturfs' (14%),'fake followers' bots purchased to increase follower counts (7%), 'self declared' bots from botwiki.org (5%), 'spammers' (2%) and 'financial bots' (0.4%). Although 'miscellaneous other bots' are quite well represented within the XR discourse online, only 17% of human users had direct interaction with them. As implied by the topic of our analysis, the individuals in our sample primarily engaged with 'astroturfs' (38%). These automated accounts are specifically designed to participate in political discussions[2], leading us to analyze users interacting with astroturfs separately from those who interacted with non-astroturfs.

Figure 4 visualizes the predicted daily change in the average number of daily tweet counts (Panel A), and the sentiment of the tweets (Panel B) grouped by different bot probabilities (65,70,75) and subbot category (astroturf or other type of bots). We found that interaction with astroturf bots results in an increase in the number of tweets, while communication with other kind of bots result in a decrease in the number of tweets. This suggests that the most politically relevant bot category—astroturf bots—drive the conversation by provoking engagement from human users, while other kinds of bots have a rather negative, silencing effect. Regardless of the bot type, direct interaction with a bot decreases the average sentiment of human users (See Tables S22–S27 for DID models).

Astroturfs tend to be mobilized in a targeted way against users with a specific opinion[15,55]. Therefore, we classified bot-exposed and matched human users by their support of XR Protests: Supporters (Pro—52%), Neutral (27%), and Anti-XR (Con—21%) using ChatGPT. (See Materials and Methods on Support Categorization). Figure 4, panel C, D indicates that human users who support or have a neutral opinion about XR are significantly affected by interacting with bots, while anti-XR users are not affected. Bot interaction has the strongest negative impact on the sentiment of bot-exposed users with neutral opinion, indicating that bots might target those users whose opinion can be changed[15,56]. The change in the number of tweets was significantly increased by bot interaction for XR supporters within the least selective bot probability category (65), although the trends are similar in the more selective categories (.70, .75). In these models, we control for the astroturf score of the interacted bot, which has a strong positive significant relationship with the change in amount. Our results indicate that the type of bot matters more for the activity change than the original support towards protests. However, the level of sentiment change depends on the original support level of users exposed to bots (See SI Table S28–S33 for DID models).

We also investigated whether bot-affected users alter their levels of support as a result of bot interaction. There were no significant differences between the distribution of bot-exposed users' support level before, and after interacting with a bot (Mann-Whitney U = 45536.00, $p = 0.86$).There was no significant difference either between users exposed to bots and those matched in terms of the change in support level.(See Model Tables S34–S35 in SI on Opinion Change). Out of the users who interacted with bots, only 9% experienced a shift

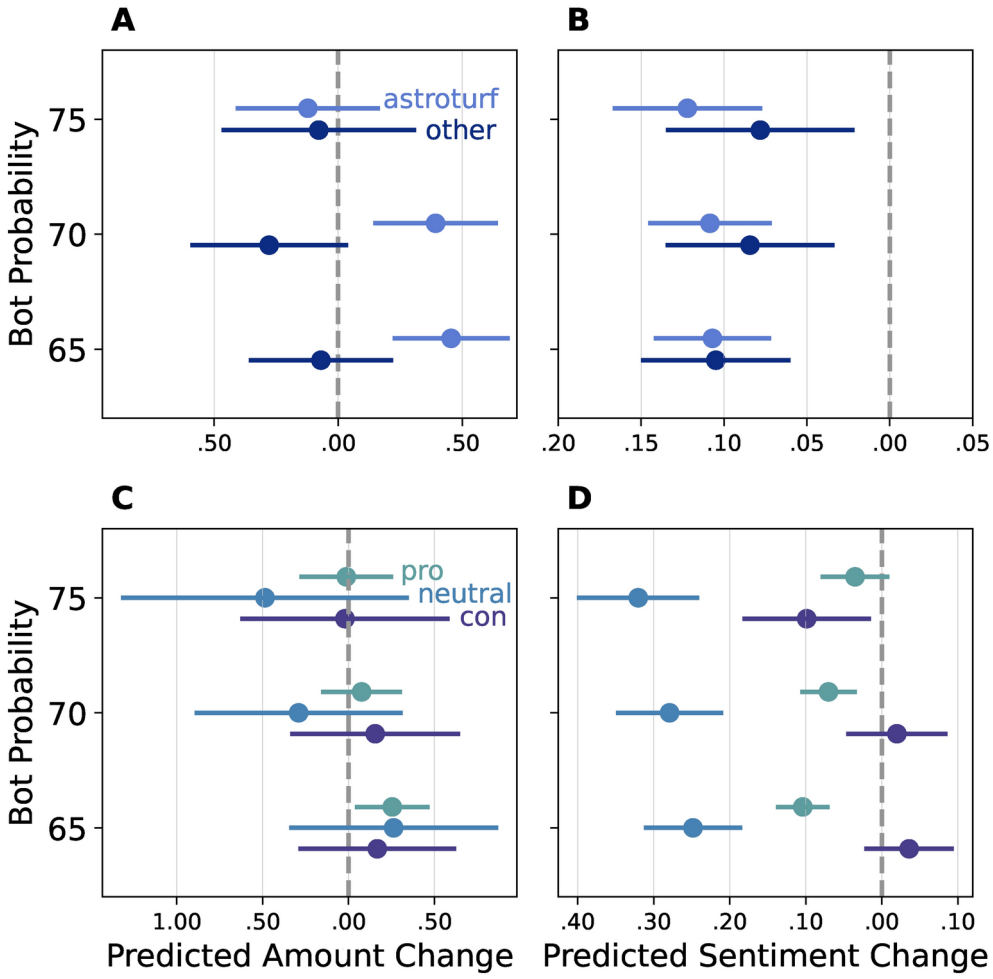

Fig. 4. Predicted change resulting from a bot interaction. Panel (**A**, **C**): Predicted change in Amount (number of tweets). Panel (**B**, **D**): Predicted change in Sentiment. Panel (**A**, **B**) separate predictions (by color) are shown for astroturf bots (light blue), and other bots (dark blue), while Panel (**C**, **D**) separate predictions by users support level towards XR—supporters (green), neutral users (blue), anti-XR (purple).

towards neutrality, while 10% of the opinions of the users in the matched group became more neutral towards XR protests. Additionally, 7% of the users exposed to bots and 6% of the matched users changed their opinion completely, shifting from negative to positive or the other way around.

The effect of bot interaction on human behaviors remains statistically significant after controlling demographic variables (location of users), for retweets to news media reports, and with a sample matched with different matching methods (See Model Tables S36–S39 in SI on other robustness checks).

Limitations

There are four potential limitations to our current research design. The first is the relatively short time range and sample size of our dataset on XR protests. Our dataset covers several waves of XR protests within a month, but XR is a global phenomenon that has been going on for several years. However, comparing our sample size to previous studies on online activism and political communications on Twitter[34,36], we believe that it is a valid and representative case to illustrate the studied aspects of bot activities on online activism. The Twitter academic product track API (available at the time of data collection) provided the full archive of tweets based on specific search queries; therefore, our dataset is a comprehensive sample of the online record of bot and human activities during the protest period.

The second limitation concerns our binary bot detection method. It has three main drawbacks, each of which we address through additional efforts and measures. First, similar to other bot identification approaches[10,48,57,58], and due to the nature of unsupervised learning, our method cannot be 100% sure whether a user classified as a bot is genuinely a bot. Second, the concept of "bot" encompasses varying degrees of automation among Twitter users, and using a fixed bot classification threshold overlooks these nuances. We are aware that being a bot, similar to the concept of gender[59], is not a binary classification problem. Many human users apply automation to increase their efficiency[60], which does not turn them into bots, but they are no longer non-automated human users. Third, a fixed threshold-based approach can still lead to false positives and false negatives, potentially undermining the validity of our causal inferences regarding bot activities. To address those concerns, we ran our models with varied CAP scores. We also performed a series of robustness checks involving various thresholds for all of our analysis and reported them in the main text. (See Discussions, SI Information flow, Cascades, and Difference-in-Difference regression results).

The third category of limitations is due to the automatized methods used to label users' support towards XR protests and sentiment of the collected tweets. We are aware that these methods are not perfect[61]. Therefore, we validated these results by two manual coders based on a subset of tweets and users timelines. We found that ChatGPT produced support level values correlated highly with human coders ($corr = 0.88$). Based on assessments of the accuracy of the VADER algorithm used for sentiment analysis, we found that it categorizes slightly more tweets as neutrals than human coders did. However, we found that the results are sufficiently accurate in all sentiment categories. (See Precision, Recall and F-score by sentiment category in SI, Table S21)

The final limitation concerns our Difference-in-Difference sampling design. We collected a comprehensive archive of tweets during the sampling period to ensure that our control (matched) group did not interact with social bots during that time. However, we cannot definitively confirm whether matched users had interacted with bots prior to the sampling period. Additionally, while we accounted for several confounders, unmeasured factors may still influence the observed differences between the treatment and control groups. Such limitation introduces a potential issue for our causal inference, and necessitate careful interpretation of the relationships.

However, our findings show that orchestrated bot activity was highly concentrated and bursty around key protest-related events, suggesting that bot influence was reactive to these news peaks. It is unlikely there was much sustained bot activity before the outburst of protest events, which implies that any potential effects from earlier interactions would be minimal. Furthermore, based on our observation of changes in sentiment and retweeting behavior, the effects of bot interaction appear to be short-lived. As a result, any bot interaction that occurred prior to our sampling period would likely have diminished by the time our analysis begins. Therefore, it is unlikely that pre-sampling bot interactions would significantly impact our estimates.

Furthermore, our analysis already reveals significant differences in behavior between users who interacted with bots and those who did not during the sampling period. If some control group users had interacted with bots before the sampling period,we would expect them to display similar behavioral changes, like shifts in sentiment. In this case, both groups would have users who have experienced bot-induced behavior changes. As a result, the observed differences between the two groups would become smaller than the actual effect, making our estimates of the treatment effect more conservative.

Discussion

Bot presence is considerable and possibly increasing in the public sphere. Even with stricter thresholds, bot activity is higher in our sample than in previous related work published on the 2017 Catalan referendum by Stella et al.[16], where only 19% of all interactions were from bots to humans. It was shown that tweets from conservative bots are more retweeted by humans, which can indicate that this difference is not only due to the continuously increasing presence of bots[62], but could be due to the highly politicized and international nature of our context.

Even though we considered only one public in detail, we replicated our descriptive analysis of information flows on similar data from Twitter about #BlackLivesMatter movement. At the active phase of exchanges from the time of the first protest followed George Floyd death, we found that the patterns of how bots impact human communication are are not significantly different in the two context (See SI Figure S1 for more details). Although XR is one of the most prominent fields of protests online, with highly committed activists which can lead to

increased bot presence and more targeted actions against them, the similarity between the two cases indicate that increased bot presence is a general threat that activists face in social media.

We have adopted a dynamic approach to political communication and have found that bots should not be thought of a constant presence, but rather bots are driving the amount of heated and bursty discussions, and react in a dynamic fashion to human sentiment. Our temporal analysis of human and bot activity showed that in 2 out of 4 identified cascades bots impacted human tweeting activity significantly more, however humans drove the sentiment of the communication. We also found that bots and humans can mutually drive cascades, depending on the topics and the intensity of the debate.

To quantify the causal impact of human–bot communication, we have compared the communication histories of human users who have directly communicated with a bot with those matched human users who have not. This allowed us to see a consistently negative impact of any kind of bot interaction on the sentiment of subsequent human communications: Humans who have interacted with a bot displayed considerably more negative sentiment than matched users.

The change in tweeting activity after a bot interaction depends on the nature of the bot: On one hand, decidedly political astroturfing bots (aiming to influence public opinion behind an impression of grassroots opinion) result in an increased activity. On the other hand, interacting with other kinds of bots (spammers, fake followers) results in a decrease in activity. However, change in sentiment towards the protests depends on the original support level of the user, supporting, neutral, or against XR. Our results indicate that bots might target users whose opinions are easier to change, since the sentiment of bot-exposed users with neutral opinion decreased the most. However, bots do not make human users switch their support level. In sum, bot interaction is not without impact, even if one encounter itself has only a small effect (it takes about two bot encounters to induce one additional tweet). Nevertheless, since there is an exceeding amount of bot communication, these small impacts add up to influence the public sphere in a profound way.

Although our analysis covers a period of time prior to the launch of ChatGPT (2022.11.30), it is becoming increasingly difficult to identify bots due to the rapid advancement mimicking human behavior[63]. As large language models are becoming widespread and easily accessible through APIs, new social bots can act extremely human, making it currently almost impossible to distinguish between bots and humans, even for experts[64,65].

Therefore, it is crucial to have unrestricted access to social media data to assess the influence and prevalence of these new types of social bots, although recent trends show that social media platforms are less willing to share free data for research purposes[66,67]. Since financial evaluations are highly correlated with the size of the (human) user base[68], platforms have no interest in quantifying the ratio of non-human accounts and their impact on misinformation[15]. Most users still underestimate the effect of bots on themselves, but as they are exposed to increased bot presence, they tend to prefer stricter bot-regulation policies[69]. That is why we welcome the news that the European Union requires larger platforms to provide researchers with access to data to study systemic risks arising from the use of their services, such as disinformation[70]. Such legislative actions can help the scientific community continue its work to understand the consequences of this abrupt change in technology that will alter the nature of human–bot interactions[64,71–73].

Data and methods

Data

Our data is made up of Twitter activity around several waves of Extinction Rebellion climate change protests from 18 November 2019 to 10 December 2019. The dataset was collected from November to December 2022 via the Academic Research product track API provided by Twitter[74], which enabled users to collect a full archive sample of historical tweets filtered based on keywords and conditions. We collected all tweets posted during this period of time that contained the keyword 'Extinction Rebellion', 'climate change protest', 'XRebellion', 'XR' and multiple variants of keywords with slightly different spelling. (The complete list of keywords used can be found in SI Table S2). In total, the final data set contained 201,010 tweets and 122,130 users.

Bot identification

To identify social bots on Twitter, we used a combination of two sets of bot identification methods. The first is a popular Twitter bot identification tool known as the 'botometer' (formerly BotorNot), which was primarily used as a benchmark to compare other methods. The second is a set of our self-trained bot identification model trained with open source data of bots and humans to train supervised machine learning models.

Botometer is a publicly available tool that relies on machine learning. It is designed to calculate a score where low scores indicate likely human accounts, while high scores suggest likely bot accounts[75]. The algorithm considers more than 1000 features related to user profiles, friends, network structure, and activity patterns, among others. Another part of our bot identification pipeline comes from self-trained models. Training sets were derived from existing open-source data from Twitter accounts identified as 'bots' and 'humans'.

We trained bot identification models with 70% training and 30% of testing set with five types of algorithms: random forest(RF), support vector machine(SVM), logistic regression(LOG), XGboost classification (XGB) and deep learning (DL). We developed two versions of our models with ten and twenty features that were proved to be most effective for bot identification by previous studies[10,76]. The evaluation of the models demonstrated that the RF, DL, and XGB models with 20 traits surpassed other models in terms of sensitivity (true positive rate), balanced precision, precision, and F1 score. Additionally, these models exhibited strong performance in an independent test carried out on a dataset consisting of influential bots and human mimics that were active during the 2018 midterm election of the United States.

In our final bot identification approach, we combined the results derived from both sets of bot identification methods. Due to the potential false positive issues inherent in both methods, they yielded somewhat divergent results[49]. To reconcile this disparity, we classified users based on the overlapping results of botometer *and* our

proprietary algorithms (DL, RF, or XGB). If both the botometer and at least one of our algorithms identified a user as a bot, the user was classified as such; conversely, the same principle was applied to the classification of humans. To account for the potential error in bot identification, we performed all our analysis with a varying baseline CAP score ranging between 0.60 and 0.75 (See more details in SI Bot Identification).

Topic modeling

We first identified the themes emerging in the protest-related discourse with the increasingly popular bi-term topic models[50,77] that learn topics by modeling word-word co-occurrence patterns[78]. After removing all retweets, to preprocess the data we used the nltk python package[79] to remove stopwords, usernames, emojis and links from the tweets, and lemmtize and stemm every word.

We then trained bi-term topic models on our preprocessed data with the bitermplus package[80]. We set up a biterm topic model of all the tweets related to XR, which classified all tweets into 8 different topics. Based on the u_mass coherence scores[81] (See Figure S5 in SI), we determined that 8 topics fits our dataset the best. After evaluating the meaning of topics, we dropped one topic whose keywords and content were too diverse to extract a meaningful media agenda from it. (See more details in SI Topic Modeling).

Testing time-series relations

To test whether bots influence human activity or vica versa we employed Vector Error Correction Models (VECM). VECM can handle multivariate time series data that are likely cointegrated, which means that they share a stable long-term relationship despite short-term fluctuations[53]. Therefore it is more suitable compared to other methods, such as vector autoregressive (VAR) models, or granger causality tests it does not require stationary time data.

We created aggregated time series of the number of posts by bots and humans, then introduced a 30-minute time-lag between the number of tweets and the average sentiment by bots and by humans. The augmented Dickey-Fuller (ADF) test, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests and the Johansen test all suggested that all time series we tested were non-stationary (e.g., the mean or variant was not constant or there were seasonal fluctuations in the time series trend), but stationary for their first order difference. However, first-order differences of the time series were stationary, suggesting that they satisfied the prerequisite of VECM. The formula of VECM models is as follows:

$$\Delta Y_t = \alpha + \sum_{i=1}^{p-1} \Phi_i \Delta Y_{t-i} + \sum_{i=0}^{q} \Gamma_i \Delta Z_{t-i} + \beta_1 (Y_{t-1} - \beta X_{t-1}) + \epsilon_t$$

where ΔY_t represents the vector of first differences of m variables $Y_1, Y_2, \ldots, Y_m$ at time t. α stands for the constant term, Φ_i are coefficient matrices for the lagged first differences of Y variables up to $p-1$ lags, Γ_i are coefficient matrices for the lagged first differences of Z control variables up to q lags, β_i is a coefficient vector capturing the speed of adjustment (error correction term), Y_{t-1} denotes the lagged level of the Y variables, X_{t-1} denotes the lagged level of the X control variables, and ϵ_t represents the error term or residual at time t.

Apart from our main variables of interest (bot posts and human posts), we also included control variables in our VECM models. We control for the number of retweets of major news media communication relevant for our geographical coverage (US, UK, Canada): BBC News, The Guardian, The Telegraph, The Independent, Sky News, Channel 4 News, ITV News, Daily Mail, Financial Times, Metro Uk and CNN. It is primarily news media reports that prompt activists and social bots to react and engage, thus events of news media communications could be the third variable that influences the timing of both human and bot engagement.

If the number of posts by bots could predict (statistically significant coefficients in VECM models) that of humans, then we would infer that bots were directing humans' attention, and vice versa. If no statistical significant was observed on both sides, or if the effect was mutual, following existing research practice, we concluded that bots and humans were driving the cascade together.[82–84] (See SI Cascades, Table S1 for detailed information of the VECM test results on all identified cascades, and further statistical robustness tests.).

Sentiment analysis

For sentiment analysis, the VADER package was used, which is an open source rule-based model and has been proven particularly effective for the classification of feelings in text from social media[85]. We used the raw text of tweets for sentiment analysis, as suggested by the package documentation. For each tweet, the algorithm assigns a sentiment score from -1 to $+1$, with -1 being the most negative, $+1$ the most positive, and 0 neutral. (See SI Figure S6 and S7 for the sentiment distribution of all tweets by exposed and matched users, and Section "Validating Sentiment Analysis", SI table S21 for details on the accuracy of the VADER python package.)

Matched sample

In order to understand how bots shape in the longer term (30 days after bot interaction) human tweeting activity and tweet sentiment related to XR, we created a sample of matched human users who did not interact with bots in our dataset to compare with 'bot-exposed' human users. We define bot interaction as very direct and active engagement of retweeting a bot (such as replying to a bot's post), not merely seeing a post from a bot. We first identified a group of 506 users in our dataset as our exposed sample, the ones who directly interacted with bots by quoting or replying to bots. Then we calculated a similarity metric between all the 'non-exposed' human users and exposed users. Specifically, we calculated Eucledian distance based on the following metrics: statuses count, followers count, friends count, favorites count, listed count, followers growth (average number of

followers increased on a daily basis), friends growth, favorites growth, listed growth, follower friend ratio. These traits were selected and/or calculated based on the user profile collected via Twitter API V1.2. The formula for the Euclidian distance is as follows, in which p_n and q_n means the N_th trait for the sample and the matching:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$

It is worth mentioning that not all exposed users have matched users, and not all matched users were active during the time window in discussions related to XR. This is because the distribution of activity levels in online political communication is right skewed[34]; half of the users posted less than five times about XR in our data set. After dropping non-active potentially matched users, in total we had 184 matched users for 303 exposed users. If a user has more than one matched user, we include up to the top five matched users in our dataset.

Our grouping passed the parallel trend assumption test (Chi-squared, $p = 0.076$ for user sentiment, and $p = 0.967$ for amount of posts), indicating no significant difference in the slopes of the trends for the treated and control groups before the treatment. We also applied Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM) which yield similar categories. (See SI Matching Design for details on matching-related robustness checks.)

Identifying cascades

Cascades were identified by calculating the temporal density (the percentage of tweets that belong to a given topic in a given time period) of each topic. Based on previous studies on the life cycle of information cascades online[86,87], the time window was specified as 1 hour.

To identify cascades, we identified bursty periods by calculating the Z-scores of the average topic density per hour for each topic, adopting methods by similar studies in social media[86,87]. We then filtered out all time units that had a Z-score larger than 2 (> 95% percentile) in any two or more consecutively two one hour time windows. Because the Z-score and the topic density could be high in time windows with only a few tweets, we also dropped those topics with no more than 50 tweets in at least one hourly time window. (See SI Cascades for detailed information on all identified cascades.)

Support group categorization using ChatGPT

For our research objectives, we also categorize users' opinions regarding the protests they engage in discussions about. To achieve this, we seek to determine whether users are in favor of or against the protests they discuss. This was measured using a scoring system ranging from −1 to 1, where −1 indicates complete opposition to the protest, 1 denotes strong support for the protest, and 0 signifies a neutral stance or unrelated discussion in their tweets. Subsequently, we employed a tri-category classification scheme for further analysis: scores between −1 and −0.1 are classified as "Anti-XR (Con protest)", those from 0.1 to 1 as "Supporters (Pro protest)," and scores between −0.1 and 0.1 as "Neutral."

The data used for this classification consists of users' tweets from our dataset. We evaluated the opinion expressed in all interactions (human replies to bots) between our sample users and bots. Additionally, for each bot-exposed user and their matched counterparts, we classified their opinions based on all tweets from their timeline before bot interaction.

We employed OpenAI's large language model (LLM), ChatGPT 3.5[88], to classify users' standpoints. This method has been raised and adopted in various studies[89,90]. For each user, we provided the model with an instruction prompt on how to classify their opinion toward climate change protests in general, along with the text to classify (users' tweets), and the model outputed the aforementioned score. To generate the three opinion scores mentioned above, we used the full timeline for each user before bot interaction (for opinion before), the full timeline after bot interaction (for opinion after), and human replies to bots' tweets during bot interaction (for opinion during interaction). (See SI Support Group Categorization, Table S4 for detailed information on prompt engineering and verification.)

Difference-in-difference models

We applied difference-in-differences (DiD) analysis to assess the effects of the two-way treatment on human users who directly engaged with bots. DiD is a statistical model design that incorporates both a treatment and a control group. In this approach, we estimated the causal effect of treatment by analyzing time series data from both treatment and control groups. We compared the treatment effect of users who had direct interaction with bots and those who did not, 30 days after direct interaction.

The estimator and formula of a DiD model is as followed[54]:

$$Y_{it} = \alpha + \beta_1 \mathrm{Treat}_i + \beta_2 \mathrm{Post}_t + \beta_3 (\mathrm{Treat}_i \times \mathrm{Post}_t) + \epsilon_{it}$$

In the estimator, Treat_i is the key explanatory variable of differences in the treatment state, and Post_t is the dummy temporal variable that says if it is before or after treatment.

For models estimating impacts on tweeting 'amount', we included days without any records of tweets (zero-tweet days) into our dataset for estimation. Because of the excessive number of zeros in the dependent variable in this case, we used zero-inflated negative binomial models to calculate the effect of bot interaction on the number of tweets. Since average daily sentiment is normally distributed, we used linear models to estimate bot impact.

We also controlled for variables that can provide alternative explanations for our findings. Our topic analysis revealed that bot activity is topic dependent and bots generate cascades, consequently influencing human

communication. Throughout this process, bots might interact with humans multiple times. Therefore, we controlled for burstiness, the topic of the interaction, and the total number of bot interactions for each user. The sentiment of the interaction and the popularity of the original tweets could impact the level of activity in a thread. Longer threads may attract more bots, and we took these factors into account as well. An analysis of the demographic features (See SI Matching Design) of the sample and the matched groups indicated no statistically significant difference in gender. However, there was a significant difference in geographical locations. Consequently, we included location categories (UK and Ireland, Europe, USA, Australia and New Zealand, other locations) as an additional control variable.

In our models that compare the impact of bot interaction by support categories, we also controlled the support level of the interacted bot and their astroturf score. (See SI Tables S22–39 for full model tables).

Data availability

Analysis code and anonymized data created for the study will be available in a persistent GitHub repository upon publication. The link to the repository is as follows: https://github.com/lindali97/bot-cascade-climate-change. Please contact L.L. (corresponding author) in case anyone wants to request the data and code from this study.

Received: 2 May 2024; Accepted: 23 September 2024
Published online: 06 November 2024

References

1. Farrell, H. The consequences of the internet for politics. *Annu. Rev. Polit. Sci.* **15**(1), 35–52. https://doi.org/10.1146/annurev-polisci-030810-110815 (2012).
2. Keller, F. B., Schoch, D., Stier, S. & Yang, J. H. Political astroturfing on twitter: how to coordinate a disinformation campaign. *Polit. Commun.* **37**(2), 256–280. https://doi.org/10.1080/10584609.2019.1661888 (2020).
3. Caren, N., Andrews, K. T. & Lu, T. Contemporary social movements in a hybrid media environment. *Annu. Rev. Sociol.* **46**, 443–465. https://doi.org/10.1146/annurev-soc-121919-054627 (2020).
4. Hepp, A. Artificial companions, social bots and work bots: communicative robots as research objects of media and communication studies. *Media Cult. Soc.* **42**(7–8), 1410–1426. https://doi.org/10.1177/0163443720916412 (2020).
5. Seering, J., Flores, J. P., Savage, S. & Hammer, J. The social roles of bots. *Proc. ACM Human Comput. Interact.* **2**(CSCW), 1–29. https://doi.org/10.1145/3274426 (2018).
6. Rahwan, I. et al. Machine behaviour. *Nature* **568**(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y (2019).
7. Assenmacher, D. et al. Demystifying social bots: on the intelligence of automated social media actors. *Social Media Soc.* https://doi.org/10.1177/2056305120939264 (2020).
8. Bastos, M. T. & Mercea, D. The Brexit Botnet and user-generated Hyperpartisan news. *Social Sci. Comput. Rev.* **37**(1), 38–54. https://doi.org/10.1177/0894439317734157 (2019).
9. Keller, T. R. & Klinger, U. Social bots in election campaigns: theoretical, empirical, and methodological implications. *Polit. Commun.* **36**(1), 171–189. https://doi.org/10.1080/10584609.2018.1526238 (2019).
10. González-Bailón, S. & De Domenico, M. Bots are less central than verified accounts during contentious political events. *Proc. Nat. Acad. Sci. U.S.A.* https://doi.org/10.1073/pnas.2013443118 (2021).
11. Woolley, S. C. & Howard, P. N. Political communication, computational Propaganda, and autonomous agents: introduction. *Int. J. Commun.* **10**, 4882–4890 (2016).
12. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammin, A. The rise of social robots. *Commun. ACM*. https://doi.org/10.1145/2818717 (2016) arXiv:1407.5225.
13. Ferrara, E. Measuring social spam and the effect of bots on information diffusion in social media. *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, 229–255 https://doi.org/10.1007/978-3-319-77332-2_13 (2018) arXiv:1708.08134v1.
14. Forelle, M.C., Howard, P.N., Monroy-Hernandez, A., & Savage, S. Political bots and the manipulation of public opinion in Venezuela. *SSRN Electronic Journal*, 1–8. https://doi.org/10.2139/ssrn.2635800 (2015) arXiv:1507.07109.
15. Shao, C. et al. The spread of low-credibility content by social bots. *Nat. Commun.* https://doi.org/10.1038/s41467-018-06930-7 (2018) (arXiv:1707.07592).
16. Stella, M., Ferrara, E. & De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Nat. Acad. Sci. U.S.A.* **115**(49), 12435–12440. https://doi.org/10.1073/pnas.1803470115 (2018) arXiv:1802.07292.
17. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Political science: Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**(6425), 374–378. https://doi.org/10.1126/science.aau2706 (2019).
18. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**(6380), 1146–1151. https://doi.org/10.1126/science.aap9559 (2018).
19. Stella, M., Cristoforetti, M. & De Domenico, M. Influence of augmented humans in online interactions during voting events. *PLoS ONE* **14**(5), 1–8. https://doi.org/10.1371/journal.pone.0214210 (2019) arXiv:1803.08086.
20. Gehl, R. W. & Bakardjieva, M. *Socialbots and their friends: digital media and the automation of sociality* 1st edn. (Routledge, USA, 2016).
21. Schäfer, F., Evert, S. & Heinrich, P. Japan's 2014 general election: political bots, right-wing internet activism, and Prime Minister Shinzō Abe's Hidden Nationalist Agenda. *Big Data* **5**(4), 294–309. https://doi.org/10.1089/big.2017.0049 (2017).
22. Hagen, L., Neely, S., Keller, T. E., Scharf, R. & Vasquez, F. E. Rise of the machines? Examining the influence of social bots on a political discussion network. *Social Sci. Comput. Rev.* https://doi.org/10.1177/0894439320908190 (2020).
23. Oliveira, E.T.C.D., Franca, F.O.D., Goya, D.H., & Penteado, C.L.D.C. The influence of retweeting robots during brazilian protests. *Proceedings of the Annual Hawaii International Conference on System Sciences 2016-March*, 2068–2076. https://doi.org/10.1109/HICSS.2016.260 (2016).
24. Köbis, N., Bonnefon, J. F. & Rahwan, I. Bad machines corrupt good morals. *Nat. Human Behav.* **5**(6), 679–685. https://doi.org/10.1038/s41562-021-01128-2 (2021).
25. Mosleh, M., Martel, C., Eckles, D. & Rand, D. G. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proc. Nat. Acad. Sci. U.S.A.* **118**(7), 2022761118. https://doi.org/10.1073/PNAS.2022761118/ASSET/1B188146-FA8C-46B6-8D96-229936460725/ASSETS/IMAGES/LARGE/PNAS.2022761118FIG02.JPG (2021).
26. Bail, C. A. et al. Exposure to opposing views on social media can increase political polarization. *Proc. Nat. Acad. Sci.* **115**(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115 (2018).
27. Ross, B. et al. Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *Eur. J. Inf. Syst.* **28**(4), 394–412. https://doi.org/10.1080/0960085X.2018.1560920 (2019).

28. Mann, C. B. Can conversing with a computer increase turnout? Mobilization using chatbot communication. *J. Exp. Polit. Sci.* **8**(1), 51–62 (2021).
29. Schuchard, R., Crooks, A. T., Stefanidis, A. & Croitoru, A. Bot stamina: Examining the influence and staying power of bots in online social networks. *Appl. Netw. Sci.* https://doi.org/10.1007/s41109-019-0164-x (2019).
30. Chadwick, A. The hybrid media system. In: Reykjavik, Iceland.[Paper Presented at ECPR General Conference, 25–27 August] (2011)
31. Stukal, D., Sanovich, S., Tucker, J. A. & Bonneau, R. For whom the bot tolls: a neural networks approach to measuring political orientation of twitter bots in Russia. *SAGE Open.* https://doi.org/10.1177/2158244019827715 (2019).
32. Murthy, D. et al. Bots and political influence: a sociotechnical investigation of social network capital. *Int. J. Commun.* **10**(June), 4952–4971 (2016).
33. Salge, C. A. D. L. & Karahanna, E. Protesting corruption on twitter: Is it a bot or is it a person?. *Acad. Manag. Discov.* **4**(1), 32–49. https://doi.org/10.5465/amd.2015.0121 (2018).
34. González-Bailón, S., Borge-Holthoefer, J. & Moreno, Y. Broadcasters and hidden influentials in online protest diffusion. *Am. Behav. Scient.* **57**(7), 943–965. https://doi.org/10.1177/0002764213479371 (2013) arXiv:1203.1868.
35. Earl, J., Maher, T.V., & Pan, J. The digital repression of social movements, protest, and activism: A synthetic review. https://doi.org/10.1126/sciadv.abl8198 (2022).
36. Freelon, D., McIlwain, C. & Clark, M. Quantifying the power and consequences of social media protest. *New Media Soc.* **20**(3), 990–1011. https://doi.org/10.1177/1461444816676646 (2018).
37. Jennings, W. & Saunders, C. Street demonstrations and the media agenda: An analysis of the dynamics of protest agenda setting. *Comparat. Polit. Stud.* **52**(13–14), 2283–2313. https://doi.org/10.1177/0010414019830736 (2019).
38. Gilani, Z., Farahbakhsh, R., Tyson, G. & Crowcroft, J. A large-scale behavioural analysis of bots and humans on twitter. *ACM Trans. Web (TWEB)* **13**(1), 1–23 (2019).
39. Shirakawa, M., Hara, T., & Maekawa, T. Never abandon minorities: Exhaustive extraction of bursty phrases on microblogs using set cover problem. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2358–2367. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1251 (2017) https://aclanthology.org/D17-1251.
40. Comito, C., Forestiero, A. & Pizzuti, C. Bursty event detection in twitter streams. *ACM Trans. Knowl. Discov. Data (TKDD)* **13**(4), 1–28 (2019).
41. Carter, N. The Politics of the Environment: Ideas, Activism, Policy, 3rd edn. Cambridge University Press, ??? (2018)
42. Bennett, W. L. & Segerberg, A. The logic of connective action: digital media and the personalization of contentious politics. *Inf. Commun. Soc.* **15**(5), 739–768. https://doi.org/10.1080/1369118X.2012.670661 (2012).
43. Marlow, T., Miller, S. & Roberts, J. T. Bots and online climate discourses: twitter discourse on president trump's announcement of U.S. withdrawal from the paris agreement. *Clim. Policy* **21**(6), 765–777. https://doi.org/10.1080/14693062.2020.1870098 (2021).
44. Chen, C. F., Shi, W., Yang, J. & Fu, H. H. Social bots' role in climate change discussion on Twitter: Measuring standpoints, topics, and interaction strategies. *Adv. Clim. Change Res.* **12**(6), 913–923. https://doi.org/10.1016/j.accre.2021.09.011 (2021).
45. Wagner, C., Mitter, S., Körner, C., Strohmaier, M., *et al.* When social bots attack: Modeling susceptibility of users in online social networks. In: # MSM, pp. 41–48 (2012).
46. Alothali, E., Zaki, N., Mohamed, E.A., & Alashwal, H. Detecting social bots on twitter: a literature review. In: 2018 International Conference on Innovations in Information Technology (IIT), pp. 175–180 (2018). IEEE
47. Martini, F., Samula, P., Keller, T. R. & Klinger, U. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data Soc.* **8**(2), 20539517211033570 (2021).
48. Yang, K., Ferrara, E., & Menczer, F. Botometer 101: Social bot practicum for computational social scientists. CoRR **abs/2201.01608** (2022) arXiv:2201.01608
49. Rauchfleisch, A. & Kaiser, J. The false positive problem of automatic bot detection in social science research. *PLOS ONE* **15**(10), 1–20. https://doi.org/10.1371/journal.pone.0241045 (2020).
50. Yan, X., Guo, J., Lan, Y., & Cheng, X. 2013-a Bitem Topic Model for Short Texts.Pdf. Www, 1445–1455 (2013)
51. Goh, K.-I., & Barabasi, A.-L. Burstiness and Memory in Complex Systems (2006)
52. Easley, B.D. Chapter 16 Information Cascades. Networks, Crowds, and Markets, 483–508 (2010)
53. Pesaran, M. H., Shin, Y. & Smith, R. J. Structural analysis of vector error correction models with exogenous i (1) variables. *J. Econom.* **97**(2), 293–343 (2000).
54. Athey, S. & Imbens, G. W. Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74**(2), 431–497 (2006).
55. Varol, O., Ferrara, E., Davis, C., Menczer, F. & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. *Proc. Int. AAAI Conf. Web Soc. Media* **11**, 280–289 (2017).
56. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017 (Icwsm), 280–289 (2017) arXiv:1703.03107
57. Yang, K., Varol, O., Davis, C.A., Ferrara, E., Flammini, A., & Menczer, F. Arming the public with AI to counter social bots. CoRR **abs/1901.00912** (2019) arXiv:1901.00912
58. Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., Feng, X., Zhang, Q., Wang, H., Liu, Y., Bai, Y., Wang, H., Cai, Z., Wang, Y., Zheng, L., Ma, Z., Li, J., & Luo, M. TwiBot-22: Towards graph-based twitter bot detection. Adv. Neural Inf. Process. Syst. (2022) arXiv:2206.04564
59. Vedres, B. & Vasarhelyi, O. Gendered behavior as a disadvantage in open source software development. *EPJ Data Sci.* **8**(1), 25 (2019).
60. Chu, Z., Gianvecchio, S., Wang, H. & Jajodia, S. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?. *IEEE Trans. Depend. Secure Comput.* **9**(6), 811–824. https://doi.org/10.1109/TDSC.2012.75 (2012).
61. Van Atteveldt, W., Velden, M. A. & Boukes, M. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Commun. Methods Meas.* **15**(2), 121–140 (2021).
62. Luceri, L., Deb, A., Badawy, A., & Ferrara, E. Red bots do it better: Comparative analysis of social bot partisan behavior. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 1007–1012 (2019)
63. Mei, Q., Xie, Y., Yuan, W. & Jackson, M. O. A turing test of whether ai chatbots are behaviorally similar to humans. *Proc. Nat. Acad. Sci.* **121**(9), 2313925121 (2024).
64. Jo, A. The promise and peril of generative AI. *Nature* **614**(1), 214–216 (2023).
65. Yang, K.-C., & Menczer, F. Anatomy of an ai-powered malicious social botnet. arXiv preprint arXiv:2307.16336 (2023)
66. The Verge: Twitter's new Elon Musk API policy is already chilling academic research. https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research (2023)
67. The Verge: Reddit's API changes are already starting to shut down some apps. https://www.theverge.com/2023/6/9/23755640/reddit-api-changes-apps-apollo-shut-down-ama-spez-steve-huffman (2023)
68. TechCrunch: IRL shuts down, citing issues with fake users. https://techcrunch.com/2023/06/26/irl-shut-down-fake-users/ (2023)
69. Yan, H. Y., Yang, K.-C., Shanahan, J. & Menczer, F. Exposure to social bots amplifies perceptual biases and regulation propensity. *Sci. Rep.* **13**(1), 20707 (2023).

70. European Commission: Digital Services Act (DSA) overview. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en#:~:text=Digital%20Services%20Act%20(DSA)%20overview,online%20travel%20and%20accommodation%20platforms. (2023)
71. Solaiman, I. The gradient of generative ai release: Methods and considerations. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23, pp. 111–122. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3593013.3593981 (2023).
72. Li, S., Yang, J., & Zhao, K. Are you in a masquerade? exploring the behavior and impact of large language model driven social bots in online social networks. arXiv preprint arXiv:2307.10337 (2023)
73. Rio-Chanona, M., Laurentsyeva, N., & Wachs, J. Are large language models a threat to digital public goods? evidence from activity on stack overflow. arXiv preprint arXiv:2307.07367 (2023).
74. Twitter: Twitter API for Academic Research | Products | Twitter Developer Platform (2022). https://developer.twitter.com/en/products/twitter-api/academic-research Accessed 2023-03-25
75. Sayyadiharikandeh, M., Varol, O., Yang, K., Flammini, A., & Menczer, F. Detection of novel social bots by ensembles of specialized classifiers. CoRR **abs/2006.06867** (2020) arXiv:2006.06867.
76. Yang, K.-C., Varol, O., Hui, P.-M. & Menczer, F. Scalable and generalizable social bot detection through data selection. *Proc. AAAI Conf. Artif. Intell.* **34**(01), 1096–1103. https://doi.org/10.1609/aaai.v34i01.5460 (2020) arXiv:1911.09179.
77. Shi, L., Cheng, G., Xie, S. R. & Xie, G. A word embedding topic model for topic detection and summary in social networks. *Meas. Control (UK)* **52**(9–10), 1289–1298. https://doi.org/10.1177/0020294019865750 (2019).
78. Wijffels, J. BTM: Biterm Topic Models for Short Text. (2023). R package version 0.3.7. https://CRAN.R-project.org/package=BTM
79. Bird, S., Klein, E., & Loper, E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. " O'Reilly Media, Inc.", ??? (2009).
80. Terpilovskii, M., & Kälin, C. Bitermplus package (2022). https://bitermplus.readthedocs.io/en/latest/index.htmlhttps://bitermplus.readthedocs.io/en/latest/install.html
81. Röder, M., Both, A., & Hinneburg, A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. WSDM '15, pp. 399–408. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2684822.2685324 (2015).
82. Bahadori, M.T., & Liu, Y. An examination of practical granger causality inference. Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013, 467–475. https://doi.org/10.1137/1.9781611972832.52 (2013).
83. Ceron, A., Curini, L. & Iacus, S. M. First- and second-level agenda setting in the Twittersphere: An application to the Italian political debate. *J. Inf. Technol. Polit.* **13**(2), 159–174. https://doi.org/10.1080/19331681.2016.1160266 (2016).
84. Bastos, M. T., Mercea, D. & Charpentier, A. Tents, tweets, and events: The interplay between ongoing protests and social media. *J. Commun.* **65**(2), 320–350. https://doi.org/10.1111/jcom.12145 (2015).
85. Hutto, C.J., & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Adar, E., Resnick, P., Choudhury, M.D., Hogan, B., Oh, A.H. (eds.) ICWSM. The AAAI Press, ??? (2014). http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html
86. Xu, P. et al. Visual analysis of topic competition on social media. *IEEE Trans. Vis. Comput. Graph.* **19**(12), 2012–2021. https://doi.org/10.1109/TVCG.2013.221 (2013).
87. Vicario, M. D. et al. The spreading of misinformation online. *Proc. Nat. Acad. Sci. U.S.A.* **113**(3), 554–559. https://doi.org/10.1073/pnas.1517441113 (2016).
88. OpenAI: ChatGPT: A large language model trained by OpenAI. https://openai.com/chatgpt. Accessed on March 13, 2024 (2024).
89. Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., & Tyson, G. Can chatgpt reproduce human-generated labels? a study of social computing tasks. arXiv preprint arXiv:2304.10145 (2023)
90. Wu, L., Chen, Y., Yang, J., Shi, G., Qi, X., & Deng, S. Event-centric opinion mining via in-context learning with chatgpt. In: China Conference on Knowledge Graph and Semantic Computing, pp. 83–94 (2023). Springer.

Author contributions

L.L., O.V., and B.V. designed research; L.L., O.V., and B.V. performed research; L.L. O.V. and B.V. analyzed data; and L.L., O.V., and B.V. wrote the paper.

Funding

The authors would like to thank the generous support from the MacAuthur Family Foundation and the Oxford Internet Institute - Dieter Schwarz Foundation who made this research possible. OV was funded by the European Union under Horizon EU project LearnData, 101086712.

Competing interests

Authors have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The project has obtained ethical approval from the Social Sciences and Humanities Interdivisional Research Ethics Committee (IDREC) at the University of Oxford. (Research Ethics Approval Ref Number: SSH_OII_CIA_21_109)

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-74032-0.

Correspondence and requests for materials should be addressed to L.L.