How deceptive online networks reached millions in the US 2020 elections

Ruth E. Appel[1], Young Mie Kim[2], Jennifer Pan[1*], Yiqing Xu[3], Ben Nimmo[4], Daniel Robert Thomas[4], Hunt Allcott[5], Pablo Barberá[4], Taylor Brown[4], Adriana Crespo-Tenorio[4], Drew Dimmery[4,6], Deen Freelon[7], Matthew Gentzkow[8], Sandra González-Bailón[7], Andrew M. Guess[9], Shanto Iyengar[3], David Lazer[10,11], Neil Malhotra[12], Devra Moehler[4], Brendan Nyhan[13], Jaime Settle[14], Emily Thorson[15], Rebekah Tromble[16], Carlos Velasco Rivera[4], Arjun Wilkins[4], Magdalena Wojcieszak[17,18], Beixian Xiong[4], Chad Kiewiet de Jonge[4], Annie Franco[4], Winter Mason[4], Natalie Jomini Stroud[19†], Joshua A. Tucker[20†]

[1*]Department of Communication, Stanford University.
[2]School of Journalism and Mass Communication, University of Wisconsin-Madison.
[3]Department of Political Science, Stanford University.
[4]Meta.
[5]Stanford Doerr School of Sustainability, Stanford University.
[6]Data Science Lab, Hertie School.
[7]Annenberg School for Communication, University of Pennsylvania.
[8]Department of Economics, Stanford University.
[9]Department of Politics and School of Public and International Affairs, Princeton University.
[10]Network Science Institute and Department of Political Science and Khoury College of Computer Sciences, Northeastern University.
[11]Institute of Quantitative Social Science, Harvard University.
[12]Graduate School of Business, Stanford University.
[13]Department of Government, Dartmouth College.
[14]Department of Government, William & Mary.

[15]Department of Political Science, Syracuse University.
[16]School of Media and Public Affairs and Institute for Data, Democracy, and Politics, The George Washington University.
[17]Department of Communication, University of California, Davis.
[18]Center for Excellence in Social Science, University of Warsaw.
[19]Moody College of Communication and Center for Media Engagement, University of Texas at Austin.
[20]Wilf Family Department of Politics and Center for Social Media and Politics, New York University.

*Corresponding author(s). E-mail(s): jp1@stanford.edu;
†Co-last authors.

Abstract

Deceptive online networks are coordinated efforts that use identity deception to pursue strategic political or financial goals. During the US 2020 elections, these networks reached at least 37 million Facebook and 3 million Instagram users, representing 15% and 2% of the platforms' active US adult users, respectively. Only 3 networks out of 49—1 network with explicitly political aims and 2 that appeared to use politics as a lure for profit—were responsible for over 70% of users reached. Notably, accounts unaffiliated with the networks played an important role in facilitating this reach by resharing content the three networks produced. Deceptive networks, regardless of whether their goals were political or financial, reached users who were older, more conservative, more frequently exposed to content from untrustworthy sources, and spent more time on Facebook.

Keywords: influence operations, disinformation campaigns, elections, social media

Main

In the summer of 2020, Rally Forge, a US marketing firm acting on behalf of the nonprofit Turning Point USA, operated a network of Facebook and Instagram accounts with fake profiles that impersonated US citizens from a wide range of demographic and ideological backgrounds, making thousands of posts pretending to express authentic opinions (see the report on CIB9 linked in Extended Data (ED) Table 1, also [1, 2]).

Posts from Rally Forge were reshared millions of times by Facebook users who were unaffiliated with this deceptive online network.

From June to October 2020, a group from Kosovo used a compromised comedy Page and six fake Facebook user accounts to post memes and links to off-platform websites that copied content from Fox News. These posts were reshared by Facebook users unaffiliated with the Kosovo group hundreds of thousands of times (see the description of FMO27 in ED Table 1).

Posting political opinions and sharing links are activities people do every day on social media. But when these actions are taken by *deceptive online networks*—which employ various types of identity deception such as using fake names and photos, deploying seemingly unrelated and independent shell accounts, and artificially boosting engagement—they violate democratic norms and can erode trust in democratic deliberation by contaminating public discourse and misleading audiences, thus potentially distorting collective preferences.

A large and growing body of research on influence operations and disinformation campaigns has focused primarily on networks pursuing political goals [3–6], particularly those originating from Russia active around the 2016 US elections [7–14]. Furthermore, these studies have focused on users' engagement with deceptive online networks rather than exposure. This limitation arises because data on when users are exposed to network content is available only to platforms, and not to external researchers. Except for one study that captured targeted ad exposure [14], previous research has inferred exposure through proxies, such as users following or mentioning network accounts, without directly measuring whether individuals viewed the content [4, 10]. For these two reasons, previous studies are fundamentally limited in their ability to explain whether and how deceptive networks reach audiences.

This pre-registered study uses platform-level aggregated data about deceptive online networks identified by Meta that targeted US-based users in the context of

the 2020 US elections to analyze their characteristics, activities, and reach. The data cover deceptive online networks that engage in political discourse around the elections, whether they pursue political or financial goals. The data we analyze encompasses 49 networks, including details on how many US adult Facebook and Instagram users saw content from deceptive online networks and how network content reached these users.

We find that deceptive networks are diverse in their origins, the number and type of accounts they operate, and their level of posting activity. The main commonality of the networks we study is the preponderance of political content they produce, regardless of whether their primary motive is political or financial. These networks reached at least 37 million unique Facebook and 3 million unique Instagram users, representing 15% and 2% of Facebook and Instagram US adult active users, respectively. However, only a few networks accounted for most of the reach—over 70% of the users who viewed any content from deceptive online networks were exposed through three networks highly active on Facebook. In addition, we find that the way these three networks reached users on Facebook was through the participation of accounts that were unaffiliated with the networks resharing deceptive network content. Regardless of the network's primary motive or how content reached users (whether directly through network accounts or indirectly through accounts unaffiliated with networks), those exposed to deceptive network content were relatively older, more conservative, more frequently exposed to content from untrustworthy sources, and spent more time on Facebook.

Deceptive online networks

We introduce the concept of deceptive online networks and define it as coordinated efforts that take place online where audiences are misled about the identity of the people behind the network. Coordination means that members of a group organize among themselves to conduct a set of activities (e.g., some members of the group

administer multiple accounts or post content at the same time). Deception refers to the concealment of the coordination and/or misrepresentation of the identities of those involved in the effort—i.e., who they really are or what they want to achieve. For example, in the Rally Forge case, workers coordinated on the content and timing of the content they produced and engaged in identity deception when they operated multiple accounts with seemingly different identities, disguising their role behind these accounts and the payment they received for posting. This identity deception is not based on content; a network can engage in identity deception without posting deceptive content such as false information.

Our study focuses on deceptive online networks that engage in political discourse, regardless of their ultimate aims. Previous research has focused predominantly on deceptive online networks where the immediate and/or end goal of the effort is political influence. These efforts have been referred to as influence operations, influence campaigns, disinformation campaigns, and propaganda campaigns [3, 6, 7, 15, 16]. Much less attention has been paid to deceptive online networks employing political discourse whose *only* goal is financial gain. Networks pursuing financial gain talk about politics to monetize online activity, e.g. by enticing users to click on ads, make payments, or inadvertently share their private financial information [17–19]. Very little research has been conducted on these networks because they are often dismissed as mere spam and clickbait, deemed unworthy of public attention and data disclosure. What we know about them comes from journalistic investigations and one qualitative study of groups in North Macedonia that became well known during the 2016 US elections for taking advantage of the elections as a hot-button issue to garner public attention and ultimately generate profits from users' clicks [20–22]. To understand how deceptive online networks shape public discourse, there is a need to expand research and data disclosure beyond networks focused solely on political influence. Our concept of deceptive

online networks addresses this need. Regardless of their motive, the networks we study engage in identity deception and interact with users on politics and social issues.

This research is part of the US 2020 Facebook and Instagram Election Study (FIES), an effort between Meta and external researchers to study the political impact of Facebook and Instagram on US adults in the context of the 2020 elections (see Supplementary Information (SI) S1.1). The deceptive online networks examined in this study include two subsets of networks removed by Meta for inauthentic behavior, which covers efforts "to mislead people or Facebook about the popularity of content, the purpose of a community (i.e. groups, Pages, Events), or the identity of the people behind it" [23]. The first subset are Coordinated Inauthentic Behavior (CIB) networks Meta removed for engaging in "coordinated efforts to manipulate public debate for a strategic goal where fake accounts are central to the operation" [24], which are "particularly sophisticated forms of Inauthentic Behavior" [23]. The second subset are Financially-Motivated Operations (FMO) networks Meta removed for inauthentic behavior that were flagged by an internal, informal Meta taskforce for producing some content related to politics in the US leading up to the 2020 US elections. This operationalization captures the key elements of deceptive online networks: coordination, misrepresentation of identity, and engagement in political discourse, regardless of a network's goal. We show CIB and FMO networks separately throughout the paper to highlight their similarities and differences, allowing for an empirical assessment of the importance of studying deceptive online networks as an overarching category. For more details on network identification, inclusion criteria, and descriptions (including names of CIB network in common usage), see Extended Data and SI S2.1.

We focus on networks operating on Facebook and Instagram that targeted the US between June 26, 2020 and February 15, 2021 (the study period). Networks identified by Meta between June 26, 2020 and November 3, 2021 are included to cover networks operating during the elections but discovered afterward. This may not capture

all deceptive online networks engaged in political discourse on Facebook and Instagram. Some networks might never have been caught by Meta. Some networks and/or information related to the networks might not have been available to Meta at the time of data analysis because the data were not preserved by the company (see SI S2.1.3 for details on data availability). Takedowns under Meta's CIB subcategory are publicly disclosed, whereas other forms of IB are often not. All attributions of network characteristics and activity are provided by Meta, including the designation of accounts as either affiliated with a network (i.e., part of it) or merely connected to network accounts (e.g., through friend or followership ties) without being affiliated with the network. Any Facebook or Instagram account not identified by Meta as affiliated with a deceptive online network is referred to as a non-network account. Some non-network accounts are connected to network accounts while others are not. These data encompass the broadest set of networks engaged in political discourse and their characteristics made available for academic research.

The dataset includes the characteristics, activity and reach of these deceptive networks on Facebook and Instagram (to access, see Data Availability section). The data on network reach that we analyze are based on active users, defined as unique adult US-based Facebook or Instagram users who were active for at least one day during the study period and excluding user accounts affiliated with networks. Network content can differ in *how* it reaches users—whether users see a post directly from the network or users see a post indirectly when a non-network account reshares a post from the network. We use the terms "direct network content" and "direct network post" to refer to original posts of network-affiliated accounts and reshares made by network-affiliated accounts. We use the terms "indirect network content" and "indirect network posts" to refer to reshares of posts of network-affiliated accounts by non-network accounts. Whenever the terms "network content" and "network post" are used in this paper, this indicates the inclusion of direct and indirect content (see SI S2.0.1 for all definitions).

Deceptive networks are diverse in origin and predominantly feature political content

Meta identified 13 CIB networks and 36 FMO networks from around the world targeting US users on Facebook and Instagram during the 2020 elections (Figure 1 Panel A). CIB networks originated from Russia (5 networks), the United States (3 networks), Iran (2 networks), China, and Romania, and included two networks whose organization spanned multiple countries (for links to detailed reports on CIB networks, see ED Table 1). One transnational network, Truthmedia (CIB3—references to networks are standardized by network type (CIB or FMO) and a number, ordered from oldest to newest date of removal) was located in a dozen countries across Asia, Europe, Oceania, and North America (see Figure 1 Panel A). As shown by the green areas in Figure 1 Panel A, FMO networks were concentrated in the Balkans (16 networks from Albania, North Macedonia, Bosnia and Herzegovina, Kosovo, Romania) and South Asia (11 networks from Pakistan, Bangladesh, and India), but also included networks from Canada, the United States, Italy, Russia, Algeria, Morocco, and the Philippines (see ED Table 1 for descriptions of FMO networks; see ED Table 2 for more details on network characteristics).

Networks varied greatly in the type and number of accounts they operated and the number of direct network posts they made (Figure 1 Panel B). On Facebook, all 49 CIB and FMO networks had activity but varied in the number of accounts (from 2 to 676) and number of direct network posts (from none in the study period to 15,273). On Instagram, 11 out of 13 CIB networks and 13 out of 36 FMO networks had accounts and the volume of direct network posts varied from none in the study period to 6,646 (see ED Table 2 and ED Figure 1 for post activity over time, see SI S4.1 for more details). In total, networks operated 4,101 Facebook accounts, creating 81,729 direct network posts, and 613 Instagram accounts, creating 16,854 direct network posts. Networks varied in the number of ads they purchased, with only 12 out of 49

networks running any ads. Those 12 networks posted 1 ad per 29 direct network posts on average, but varied from 1 ad per 0.6 direct network posts (FMO28) to 1 ad per 1,042 direct network posts (CIB6). Networks paid to make posts more visible only 0.6% of the time.

Politics and social issues are the most frequent topics for both CIB and FMO networks on Facebook (Figure 1 Panel C based on the Meta Topic classifier for Facebook, see SI S2.0.2). On Facebook, politics and social issues accounted for 65% of direct network content for CIB networks, and 32% of direct network content for FMO networks (based on Civic classifier, see SI S2.0.2). On Instagram, politics and social issues are among the most frequent topics, making up 18% of direct CIB network content and 24% of direct FMO network content (based on Civic classifier, see SI S2.0.2). The prevalence of political content is unsurprising given our aim to study deceptive online networks engaged in political discourse. However, it is notable that politics and social issues were the most frequently discussed topics not only for CIB networks but also for FMO networks. While FMO networks had to produce at least some politics content to be included in the study, there was nothing in the selection process that predetermined that political content would be as frequent. Put another way, FMO networks actively produced content related to politics and social issues targeting users in the US. More generally, politics and social issues were widely debated, and networks appear to have utilized them, along with other contentious issues of the time such as as COVID-19 and news, and niche content such as hunting, in their posts (see SI S4.2 and CIB reports linked in ED Table 1 for additional details on network content).

A few networks dominate reach

Network content reached at least 36.79 million unique US adult active users on Facebook (14.63% of approximately 250 million unique US adult active users) and 2.98 million unique US adult active users on Instagram (1.85% of approximately 160 million

unique US adult active users). An additional 545,299 unique active users on Facebook and 7,189 unique active users on Instagram were reached solely via network ads, for a total reach of 37.33 million (14.85%) unique active users on Facebook and 2.98 million (1.86%) on Instagram over our 8-month study period. This is roughly equivalent to 15% and 1% of the US voting-age population, respectively (based on US government census data, see https://www.census.gov/content/dam/Census/library/publications/2022/demo/p20-585.pdf). For context, in the 2016 US elections, the potential exposure of the Russian Internet Research Agency is estimated to have been 126 million Facebook users over a 2-year period [25]. Distinguishing between exposure to CIB and FMO networks, and excluding users who were reached only via ads, 15.3 million (6.09%) Facebook users saw content from CIB networks and 27.7 million (11.01%) from FMO networks; 2.94 million (1.83%) Instagram users saw content from CIB networks and 34,570 (0.0215%) from FMO networks (the sum of those exposed to CIB and FMO networks is greater than the total number of unique active users exposed because some users saw content from both types of networks).

Those reached by the networks were often repeatedly exposed. On Facebook, in total, there were 175.24 million views of network content (excluding ads), 55.87 million for CIB networks and 119.37 million for FMO networks (see ED Table 3 for details). On average, viewers exposed to either CIB or FMO networks saw 4.76 pieces of network content, but the distribution was highly skewed such that the median is lower. On Instagram, in total there were 69.65 million views of network content (excluding ads), 69.30 million for CIB networks and 356,399 for FMO networks (see ED Table 4 for details). This means that viewers exposed to either or both CIB and FMO networks on Instagram had an average of 23.4 views of network content (see SI S4.3 for more details), but the distribution is also highly skewed with few people accounting for most exposures.

Exposure to deceptive network content accounts for a very small share of users' overall political content consumption. Prior research shows that news and politics make up less than 15% of total digital content consumption [26, 27]. During the 2020 election period, less than 10% of content users saw on Facebook was political [28]. Analyzing individual-level participant data, we find that for participants who encountered content from any deceptive online network in the 41 days before the election, such content represented, on average, 0.3% of their political content views on Facebook. The distribution was highly skewed: for users in the 99th percentile of the proportion of network content views to political content views, deceptive network content accounted for 3% of political content views.

On Instagram, 1% of active users accounted for over 96% of views of network content (see Figure 2 Panel A). On Facebook, 1% of active users accounted for 55% of views of network content (see Figure 2 Panel B). The greater concentration of exposure on Instagram aligns with the fact that viewers exposed to network content on Instagram are exposed to more content on average than viewers exposed to network content on Facebook. For context, previous research on presumed exposure on Twitter to content from the Russian Internet Research Agency in 2016 found that 1% of users accounted for 70% of exposures [10]), and research on fake news on Twitter in 2016 found that 1% of users accounted for 80% of such views [29]).

Notably, a small number of networks are responsible for reaching almost all viewers. On Instagram, nearly all viewers (95%) were reached by one network (CIB10; Instagram reach shown in yellow in Figure 2 Panel C). On Facebook, nearly 80% of all unique viewers of deceptive networks were reached by three networks (CIB9, FMO27, and FMO34; Facebook reach shown in blue in see Figure 2 Panel C). Because more Facebook users were exposed to network content than Instagram users, the three dominant networks on Facebook account for over 70% of all users who saw deceptive online network content across both platforms. It is also worth noting that many networks

have limited reach—8 networks on Facebook and 16 networks on Instagram reached fewer than 1,000 viewers (see Table S10 in SI S4.3).

Accounts unaffiliated with networks play a critical role in reach

Focusing on Facebook, where networks reached far more viewers, a clear pattern differentiating the networks with the highest reach from other networks is the participation of non-network accounts, i.e., accounts unaffiliated with the deceptive online network (green circles in Figure 3 Panel A). A Facebook user can see content *directly* posted by network accounts—for example, if the user is friends with a network user account, follows a network Page account, is recommended content from a network-affiliated account, or sees an ad from the network (green circles connected to blue circle in Figure 3 Panel A). However, if a Facebook user, Page, or Group that is not part of a network reshares an original network post, other Facebook users who see the reshare are exposed to network content *indirectly* through a non-network account (green circles connected to green circles in Figure 3 Panel A).

Figure 3 Panel B, where each column is a network and column width corresponds to the total number of viewers exposed to each network, shows that exposure through non-network accounts (green) vastly outpaces exposure through network accounts (blue), especially for networks that reached larger numbers of viewers (for details by network, see ED Table 5). The CIB network with the highest overall reach, CIB9, only reached 1.3 million viewers directly, but through the participation of non-network accounts resharing its content, CIB9 reached 13 million viewers indirectly, for a total of 13.4 million viewers reached (the sum of directly reached and indirectly reached users is larger than the total because some viewers were reached both directly and indirectly). In addition, we find positive correlations between the share of viewers reached through indirect exposure and average cascade size, average cascade depth,

average cascade maximum breadth, and average structural virality of network original posts (see SI Figure S27 in SI S4.5). It is important to note that only a small proportion of users who see network content (5.67% on Facebook and 0.34% on Instagram) reshared network content. This means that even though only a small proportion of those exposed to network content are involved in its propagation, their involvement disproportionately amplifies overall reach. In sum, during the study period, the number of Facebook users reached by network content was massively larger because non-network accounts reshared network content.

It is possible that network exposure could be affected by platform features such as the Facebook feed ranking algorithm. We conducted a post-hoc exploratory analysis of the data from FIES experiments [28, 30, 31] where consenting Facebook users were randomly assigned to (1) a feed without reshares, (2) a reverse chronological feed, or (3) a feed where like-minded content was reduced to examine how these treatments affected exposure to and engagement with deceptive online networks compared to the default feed ranking algorithm (control group). We found that the no reshares and chronological feed conditions decreased exposure to and engagement with deceptive networks. However, the treatments also decreased overall user engagement on Facebook, and due to the small number of experiment participants who were exposed to network content, we were not powered to detect the effect of these treatments after overall views and engagements are taken into account (see SI S4.6 figures and tables).

Network content viewers have similar characteristics

When we examine who is reached by deceptive online networks, we find that the characteristics of viewers are similar between CIB and FMO networks. In addition, we find that the characteristics of viewers are similar between viewers who are exposed directly by network accounts and viewers exposed indirectly through non-network

accounts. We focus on Facebook here because some characteristics such as age and high exposure to content from untrustworthy sources are not available for Instagram.

There are four panels in Figure 4, each with two plots, that show for each network with more than 100 viewers the characteristics of users exposed to network content directly through network accounts (top plot in each panel) and indirectly through non-network accounts (bottom plot in each panel). The user characteristics are political ideology (Panel A), age (Panel B), quartiles of time spent on Facebook (Panel C), and the share of users in the top 20% of Facebook users in terms of exposure to content from untrustworthy sources (Panel D). The horizontal dashed lines in each plot correspond to that characteristic in the population of US adult active users in the study period (e.g., the two dashed lines in the first pair of plots separate the share of liberal, moderate, and conservative adult active Facebook users).

For both CIB and FMO networks, those exposed are more likely to be conservative—the red portions of Panel A rise higher for the majority of networks than the lower dashed line, which shows that the proportion of conservatives exposed is higher than the proportion of conservatives in the US adult active Facebook user population. Those exposed are older than US adult active Facebook users in general—the dark and medium blue in Panel B is higher than the second lowest dashed line, which shows the proportion of US adult active Facebook users age 45 to 64 and 65 and above, for almost all networks. Both results align with previous research showing that conservatives and older adults are disproportionately reached by disinformation and misinformation [4, 10, 29], as well as research on echo chambers and selective exposure [32, 33]. Those exposed to CIB and FMO networks were also highly active on Facebook in terms of time spent, and disproportionately among the top 20% of Facebook users who are exposed to content from untrustworthy sources (for other characteristics, see SI Figures S24-S26 in SI S4.4). The characteristics of viewers who are exposed directly through network accounts (top plot in each Panel in Figure 4)

and those exposed indirectly through non-network accounts (bottom plot in each pair Panel in Figure 4) are also generally similar. Several characteristics of those viewing and those resharing content are similar and deviate from the characteristics of the general user population (see SI Figures S18-S22 in SI 4.3).

The importance of user characteristics is further highlighted by an analysis based on individual-level participant data of the relationship between exposure to and engagement with network content and downstream consequences. Based on a pre-registered analysis of 72,962 consenting survey participants, we examine the relationship between exposure to/engagement with deceptive online networks and individual-level outcomes controlling for a set of user characteristics, which were measured pre-exposure. The outcomes we examine—the ability to discern true from false information, belief in the legitimacy of the 2020 elections, and behavioral polarization (partisan news clicks on Facebook)—capture different aspects of politics (knowledge, attitudes, behavior) identified by prior research as areas potentially affected by deceptive online networks [4, 10, 29]. We set a threshold to only analyze networks whose content was viewed by 500 or more consenting participants after Wave 2 and before Wave 4 of survey data collection (i.e., between September 24, 2020 and November 3, 2020; see SI S3.1 for details). Only two FMO networks operating on Facebook (FMO27 from Kosovo and FMO35 from North Macedonia) met these criteria (for more details on these networks, see S3.2, Table S5 in SI S4.2 for network content topics, Figure S9 in SI S4.2 for candidate mentions, ED Table 1 for network descriptions).

For both FMO networks, without controlling for any individual-level characteristics, we observe that survey participants who were exposed to or engaged with the deceptive online networks tended to have less ability to discern true from false information , and were more likely to exhibit partisan news clicks (see Model 0 in the panels of Figure 5 and SI Table S3 for coefficient estimates, p-value, and confidence intervals). After adjusting for pre-exposure individual-level characteristics using

entropy balancing [34]—which reweights non-exposed/non-engaged participants to achieve mean balance on these characteristics with the exposed/engaged group (see SI S3.1.5 for more details), the previously strong associations between exposure to or engagement with deceptive online networks and users' factual knowledge (FMO27: Treatment = Exposure, $t(30,003) = -1.603$, $p = 0.109$, coef $= -0.022$, 95% CI $= [-0.048, 0.005]$; Treatment = Engagement, $t(30,003) = 1.267$, $p = 0.205$, coef $= 0.031$, 95% CI $= [-0.017, 0.080]$; FMO35: Treatment = Exposure, $t(30,003) = -1.908$, $p = 0.056$, coef $= -0.048$, 95% CI $= [-0.096, 0.001]$; Treatment = Engagement, $t(30,003) = -1.580$, $p = 0.114$, coef $= -0.064$, 95% CI $= [-0.142, 0.015]$, perceived election legitimacy (FMO27: Treatment = Exposure, $t(28,952) = 0.201$, $p = 0.840$, coef $= 0.003$, 95% CI $= [-0.022, 0.027]$; Treatment = Engagement, $t(28,952) = 1.501$, $p = 0.134$, coef $= 0.032$, 95% CI $= [-0.010, 0.074]$; FMO35: Treatment = Exposure, $t(28,952) = -1.931$, $p = 0.054$, coef $= -0.040$, 95% CI $= [-0.081, 0.001]$; Treatment = Engagement, $t(28,952) = 2.230$, $p = 0.026$, coef $= 0.067$, 95% CI $= [0.008, 0.126]$) , and subsequent online behavior (FMO27: Treatment = Exposure, $t(36,211) = 1.834$, $p = 0.067$, coef $= 0.002$, 95% CI $= [0, 0.004]$; Treatment = Engagement, $t(36,211) = 2.391$, $p = 0.017$, coef $= 0.005$, 95% CI $= [0.001, 0.009]$; FMO35: Treatment = Exposure,$t(36,211) = 0.160$, $p = 0.873$, coef $= 0$, 95% CI $= [-0.004, 0.005]$; Treatment = Engagement, $t(36,211) = -1.689$, $p = 0.091$, coef $= -0.007$, 95% CI $= [-0.014, 0.001]$) largely disappear (Model 8).

We caution against interpreting these results as causal—that exposure to/engagement with network content influences (or does not influence) downstream political knowledge, attitudes, and behavior. This analysis relies on individual-level participant data, and the relatively small number of exposed participants may not represent the broader population of users who encounter network content. More generally, due to limited knowledge about how the treatments (exposure to or engagement with the networks) are assigned, causal inference using observational data based on the

unconfoundedness assumption is inherently challenging [35, 36]. Our pre-registered sensitivity analyses show that these estimates are highly vulnerable to potential unobserved confounding and therefore cannot be credibly interpreted as causal (see SI S3.2.3 for full Sensitivity Analysis results). For example, individuals who have lower trust in media, higher prior approval of Donald Trump as president, are aged 45 or older, and identify as Republican are both more likely to click on content from FMO35 *and* more likely to doubt the legitimacy of the 2020 elections.

The observed associations between exposure to or engagement with two FMO networks and political outcomes are likely confounded by unobserved individual-level characteristics and should not be interpreted as evidence for causal effects—or the absence of such effects. Altogether, these results underscore the importance of understanding the characteristics of individuals who are exposed to and engage with network content.

Discussion

Deceptive online networks use various tactics of identity deception, violating community standards and norms on social media. Deceptive online networks raise normative concerns as they may mislead the public, contaminate political discourse, and corrode public confidence in democracy.

In the 2020 US elections, a diverse array of deceptive online networks targeted users in the United States. While prior research has focused primarily on foreign influence campaigns from Russia, we find that deceptive online networks targeting the US electorate originate domestically and from around the world. In fact, the network with the largest reach in this study, Rally Forge, was operated by US domestic actors.

The results show that all deceptive online networks producing and disseminating political content should be scrutinized, regardless of their goals. Prior work on coordinated networks, influence operations, and disinformation campaigns has largely

neglected financially-motivated actors, dismissing them as mere spam or clickbait. This has resulted in little research or oversight [20, 21]. In our study period, Meta identified three times as many financially-motivated as politically-motivated deceptive networks that engaged in political discourse. These financially-motivated networks reached more users and shared a substantial amount of political content. The characteristics of users reached by deceptive online networks are similar, regardless of networks' goals. Furthermore, in practice, it can be difficult to make clear attributions of intent (e.g., was a network pursuing political or financial goals). It is also conceivable that some networks are "repurposed" over time—for example, accounts originally used for financial gain may later be sold or leveraged for explicitly political purposes. All deceptive online networks engaged in political discourse—regardless of how they are classified—may shape election contexts and deserve scrutiny, and future research could benefit from delving more deeply into the function, structural organization, and dissemination tactics of networks. . To enable this, it is essential for platforms not only reinvest in trust and safety efforts—reversing recent trends [37, 38]—but also take proactive steps to document, retain, and publicly release data on *all* deceptive online networks engaged in political discourse.

Leveraging unique data on actual exposure, we find that 37 million active users on Facebook and 3 million active users on Instagram were exposed to some deceptive online network content in the 2020 US elections. These numbers are likely a lower bound on exposure for several reasons: not all networks may have been detected by Meta, we only analyze an 8-month study period, only adult users are included, and not all network data was preserved (see SI S2.1.1 for details on network identification, SI S2.1.3 on missing data). While reach was highly concentrated on Instagram, with 1% of users accounting for 96% of all views of network content, it was more diffuse on Facebook, where 1% of users accounted for 55% of all views. Notably, only a very small number of deceptive networks achieved broad reach, and they accounted for the vast

majority of exposures to deceptive network content. As with all of our results, care should be taken when generalizing these findings beyond the context of this study.

Reshares by non-network accounts, i.e., accounts that are unaffiliated with deceptive online networks, are crucial for exposing more users to deceptive online network content. Dovetailing with previous research, this suggests social media users are often active participants in amplifying deceptive network content, not just passive information consumers [10, 39–45].

The importance of non-network accounts for network reach has implications for future research and the development of solutions and countermeasures. First and foremost, it highlights the need for future research to focus on understanding the motivations and behavior of the relatively small number of users who choose to reshare deceptive network content. For example, these users may be influenced by social reinforcement or ideological affinities [46], or they may rely on heuristic processing and intuitive decision-making [47, 48] when they share network content. These users, particularly those directly exposed to networks, play a key role in the transmission chain, connecting the networks to other users who would not have been directly exposed to deceptive networks but are nevertheless reached indirectly through reshares of non-network accounts. Since humans judge reliability based in part on others' interactions, when deceptive network content is reshared by non-network users, it may make identity deception more difficult to detect [49, 50]. This suggests that countermeasures aimed solely at the networks themselves may be insufficient unless paired with behavioral interventions or preventative strategies that address the reasons users choose to reshare deceptive network content.

Although we illustrate the role of non-network accounts in users' exposure to deceptive networks, we do not know how exactly non-network accounts facilitate spread. Our analyses of engagement with network content, such as reshares, by different user

characteristics such as age and posting activity give insight into who reshares network content (see SI S5.1). Our data on network original post diffusion shows averages of cascade size, depth, maximum breadth, and structural virality across all diffusion trees per network (see SI S5.2 Tables S39-S42 for select diffusion results; results for all pre-registered analyses can be found in the Supplementary Repository (SR), see SI S7 and https://bit.ly/SupplementaryRepo (passcode: aZXGdq), including all diffusion results in SR2.4). However, we cannot identify whether deceptive networks intentionally targeted accounts with many followers for amplification, whether networks targeted non-network accounts with specific characteristics, or whether characteristics or behaviors of non-network accounts, beyond what is measured in this study, promote content spread [51]. This suggests a potential role that platform-specific features—for example, reshare functionality on Facebook—may play in enabling broader reach of deceptive networks, which should be explored in future research (see SI S4.6 for an exploratory analysis). The effectiveness of countermeasures would be strengthened with a deeper understanding of the mechanisms under which non-network reshares occur.

The emphasis on non-network accounts does not necessarily mean that future research can ignore deceptive networks' direct actions and reach. On Instagram, which has different reshare features from Facebook, CIB10, the most successful network in terms of reach on Instagram, reached 2.8 million active users directly. Even on Facebook, FMO29 reached 2.2 million active users directly, without the amplification by non-network accounts. Networks can directly reach millions of users and do so in a number of different ways. Networks can compromise popular accounts—for instance, FMO27 compromised a comedy Page to connect directly with Facebook users. Networks can use targeted ads to reach desired audiences. Rally Forge (CIB9), the most successful network in terms of reach on Facebook, utilized Facebook ads (513 in our

8-month study period, see ED Table 2), spending about $1.15 million according to Meta's public reporting [52] (over a potentially longer period).

Finally, it remains important to understand the causal effects of deceptive online networks, including not only whether viewers' beliefs are changed, but also whether beliefs are hardened, and how networks affect general public perceptions. In our observational data, we were unable to reach firm conclusions about the causal effects of networks. Future research could focus on sampling people likely to be exposed to networks to improve power, or run longer-term studies.

The data underlying this study, including all figures and tables in the SI and SR, is available to create opportunities to analyze deceptive online networks further (for details on access, see Data Availability). Platform transparency and data sharing with researchers is essential in further investigating the role and impact of deceptive online networks.

We aim to promote clarity in public debates and to foster evidence-based policymaking in three main ways: 1) providing a conceptualization of deceptive online networks that moves beyond networks' goals, 2) directly measuring exposure to deceptive networks, and 3) employing a variety of approaches to show how platform users are exposed to deceptive networks. The nature of deceptive online networks is complex, and the mechanisms of their reach are multi-faceted. However, this study highlights the key role of social sharing. Accounts unaffiliated with deceptive online networks played a pivotal role in spreading network content to broader audiences. Understanding the importance of resharing by non-network accounts and the characteristics of the users who inadvertently become accomplices in deception is crucial for effectively addressing the issue of deceptive online networks on social media.

Methods

Ethical considerations

We considered a number of ethical concerns related to the research and designed the studies to minimize potential harms to the respondents involved in them, as well as any broader social harms. Meta sought review from and was granted approval to conduct the experimental studies by the NORC Institutional Review Board (Protocol number 20.08.10, Project number 8870). Academic collaborators worked with their respective university IRB's to ensure compliance with Human Subjects Research regulations in their authorship of papers, including analysis of aggregated, de-identified data collected by Meta and NORC. Meta retained the services of Ethical Resolve, a data ethics firm that was consulted by both Meta and academic researchers at various stages of the project prior to implementation of the research to evaluate whether it met long-running traditions of research ethics as well as emerging norms and best practices for conducting digital research.

Definitions

We define terms in SI S2.0.1, such as what constitutes a "view" (e.g., on Facebook, a view is counted whenever the post renders in the visible portion of a user's web browser or mobile device for more than 250 milliseconds).

We describe and show the performance of classifiers as well as other categorization methods that are applied to the data, such as those used to identify political content, in SI S2.0.2 and S2.0.3.

Study design

This study was initially pre-registered on October 16, 2020 at https://doi.org/10.17605/OSF.IO/D7QES (all deviations and clarification can be found in SI S8). The study utilizes multiple types of data, including aggregated

platform data, individual-level participant data, and network data. We provide an overview of each data type here and refer to relevant SI sections for more details.

Data

Aggregated platform data: Aggregated platform data refers to data on all adult U.S.-based Facebook and Instagram users who were active for at least one day between June 26, 2020 and February 15, 2021 (the study period). The size of this sample is about 250 million adults for Facebook, and about 160 million adults for Instagram. For Facebook, about 46.8% of users are male, 53.2% are female; 26.7% are aged 18-29, 30.7% 30-44, 29.2% 45-64, and 13.3% 65 plus; 28.6% are politically liberal, 19% moderate and 50.4% conservative. For Instagram, about 42.4% of users are male and 57.6% are female. This sample is representative of the adult active Facebook and Instagram US user population, but not the general US population. For more details, see SI S2.2.

Individual-level participant data: Individual-level participant data refers to data from U.S. Facebook and Instagram users age 18 and over who consented to participate in a study of social media and politics and completed two baseline survey waves. The data include individual-level survey responses as well as participants' on-platform behavioral data. The sampling frames included all Facebook and Instagram monthly active U.S.-based users 18 years of age or older eligible to receive general surveys on a given platform (these represent a random set of users from the overall Facebook and Instagram populations) as of August 17, 2020. Participants were asked to confirm they were over 18 years of age and lived in the United States as part of the recruitment process. The total sample size is about 73,000, although we use different subsamples for different analyses. Of these users, 43.3% are male, 56.7% are female or another gender (after excluding missing values). 17.9% are aged 18-29, 41.8% 30-44, 33.3% 45-65, and 7% 65 plus (after excluding missing values). In terms of political identification, 52.4% are Democrat or Democrat-leaning, 34.9% Republican or Republican-leaning,

and 12.8% Independent. This sample is neither representative of the adult active Facebook and Instagram US user population, nor the general US population. For more details, see SI S2.3. For individual-level data, we excluded participants who deleted their account, withdrew from the study, or requested data deletion from the analysis for ethical reasons since we could not access their data.

Network data: Deceptive online networks are operationalized as networks disabled under the Coordinated Inauthentic Behavior (CIB) protocol, which consisted of influence operations, whether domestic or foreign in origin, that attempted to manipulate or corrupt public debate for a strategic goal. Network also include those disabled under the Inauthentic Behavior (IB) protocol. These consisted of financially motivated operations (FMO) that were identified by an informal, internal Meta taskforce as networks using inauthentic accounts to post content on American political themes—usually to drive users towards off-platform websites where the operators could monetize users' clicks. For more details on data related to deceptive online networks, including how networks were identified, how these data were prepared, missing data, and more detailed descriptions of networks, see SI S2.1. Note that our operationalization of network content excludes network comments, which some networks like Rally Forge commonly used [52], implying that our numbers on content production and reach are underestimates. Further, due to lower performance for other languages, we focus on posts in English and Spanish for topic classifications, meaning that we do not take posts in other languages into account when assessing the share of political content or topics about which a network posted.

Sampling strategy

Aggregated platform data: This represents the population of US-based adult active users on Facebook and Instagram. For more details, see SI S2.2.

Individual-level survey data: Below, we summarize our sampling strategy. Further details are available in SI S2.3.2. Participants were sampled to achieve specific sample targets across different stages of the study, which were in turn chosen to achieve desired minimum detectable effect sizes (MDEs) for experimental interventions related to other papers part of the US 2020 FIES across different subgroups among the set of respondents participating in the Wave 1 and Wave 2 surveys and Waves 4 or 5. The sampling frame consisted of all U.S. Facebook and Instagram users aged 18 and above who were active monthly users and qualified to participate in standard platform surveys as of August 17, 2020. This participant pool constituted a randomly selected subset of the broader user bases across both social media platforms. The Facebook sampling frame was trimmed by removing predicted fake accounts, employees, and advertisers. For Instagram, creators and business accounts also were removed. Finally, because the use of multiple accounts is common among Instagram users, the sampling frame was narrowed to include only a user's primary account (for users with multiple accounts, this is the oldest account). Having defined the sampling frames, sampling probabilities were computed to achieve specific sample distributions for the set of demographics encoded in the stratification step across each of the samples of interest. The sampling probabilities took into account (a) differential non-response across different demographics and (b) the desired sample size across the different studies. At the top of their Facebook feed, randomly selected participants saw a recruitment message asking them if they would like to share their opinion. Those clicking "Start Survey" were directed to a consent form. Participants gave their consent to participate using an IRB-approved consent form that outlined the study procedure, benefits and risks, and compensation. During recruitment, individuals were required to verify their age (18+) and U.S. residency status. The sample size is $N = 72,962$, with $N = 19,510$ participants randomized into the control group for the FIES platform interventions. Additional information regarding sampling methodology, recruitment, and response

rates can be found in SI S2.3. Data collection and analysis were not performed blind to the conditions of the experiments.

Network data: Deceptive online networks are operationalized as networks Facebook (now Meta) disabled between June 26, 2020 and November 3, 2021 (one year after election day) that targeted US users during the study period (June 26, 2020 and February 15, 2021). This resulted in 49 networks. For more details, including on identification of networks, see SI S2.1. For one analysis on the downstream effects of networks, we subset to networks that reached a pre-registered reach threshold between different survey waves, see SI S3.1.1 for details.

Data collection

The study's data collection involved a partnership between Meta and NORC, a survey research organization at the University of Chicago. While Meta handled the majority of participant recruitment and gathered information from Facebook and Instagram platforms, NORC managed the survey administration process, including expanding the participant pool through additional recruitment, gathering supplementary information beyond the social media platform data, and ensuring participant anonymity by stripping identifying information before data integration and distribution to researchers. Participants provided data through their personal devices, including smartphones and computers. NORC operated without knowledge of the research hypotheses or participants' experimental assignments.

On-platform behavioral data and network data were collected via Meta's internal systems for logging user behavior. Survey data were collected by NORC using their existing survey infrastructure. To collect the passive measurement data, NORC partnered with two vendors: MDI Global and RealityMine.

Aggregated platform data: For Facebook, aggregated data are based on computing the total count for the study period (2020-06-26 until 2021-02-15). For Instagram,

metrics are aggregated between 2020-07-17 and 2021-02-15 due to an unanticipated data gap.

Individual-level survey data: Survey data collection started on August 31, 2020. Two surveys were fielded pre-treatment: Wave 1 (August 31-September 12) and Wave 2 (September 8-23). The treatments that we analyze in more detail (less like-minded content, no reshares, reverse-chronological feed) ran from September 24-December 23, other treatments had different timelines (see SI 3.1). During the treatment period, three more surveys were administered: Wave 3 (October 9-23), Wave 4 (November 4-18), and Wave 5 (December 9-23). See SI S2.3 for details on the survey and platform behavioral data for individual-level participants and data collection period details.

Network data: Network data are measured for the study period (2020-06-26 until 2021-02-15). For more details, see SI S2.1.

Data Transparency

The study relies on data provided by Meta, which offers an important view of deceptive networks targeting U.S. adult users on Facebook and Instagram between June 26, 2020, and November 3, 2021. At the same time, this reliance raises methodological constraints. First, we can only analyze what Meta made available. Not all data from the identified networks were retained (see SI S2.1 for details on missing data), and the informal taskforce engaged in flagging FMOs may not have captured all relevant networks. Importantly, we cannot independently verify the completeness of the dataset, and as a result, there may be biases in the data (e.g., selection bias in identifying networks) that we cannot assess. Second, all attributions of network characteristics and activity come from Meta. Despite requests, Meta did not provide descriptions of all FMO networks because they did not have the resources in their investigative teams to conduct post-hoc investigations of all FMO networks. Meta provided limited explanation of how they determine whether an account is affiliated with a network

versus merely connected to it (non-network accounts). Despite repeated requests for details, all that was provided to us by Meta regarding network identification is what is presented in SI S2.1.1. These constraints should be kept in mind when interpreting the findings. Nonetheless, the dataset has advantages: it captures actual exposure to deceptive networks across the entire population of U.S. adult users on Facebook and Instagram, in contrast to user surveys or qualitative studies that rely on self-reports, presumed exposure, or limited samples. These data also provide a fine-grained view of networks and their activity over time. While the data lack the depth of ethnographic or interview-based approaches, they enable a population-level analysis of deceptive networks that has not previously been possible.

Data Availability

De-identified data from this study is stored in the Social Media Archive (SOMAR) housed by ICPSR. This is accessible for university research approved by the Institutional Review Board (IRB) related to elections, or for the purpose of validating the results of this study. ICPSR will handle and vet all applications requesting access to this data. Data access is controlled to protect the privacy of the study participants and to be consistent with the consent form signed by study participants. Requests for data can be made via SOMAR (https://socialmediaarchive.org/). The data for this paper is based on the following datasets stored in SOMAR: https://socialmediaarchive.org/record/87, https://socialmediaarchive.org/record/88, https://socialmediaarchive.org/record/89, https://socialmediaarchive.org/record/90, https://socialmediaarchive.org/record/91, https://socialmediaarchive.org/record/92, https://socialmediaarchive.org/record/93, https://socialmediaarchive.org/record/94, https://socialmediaarchive.org/record/95, https://socialmediaarchive.org/record/96, https://socialmediaarchive.org/record/97, https://socialmediaarchive.org/record/98, https://socialmediaarchive.org/record/99,

https://socialmediaarchive.org/record/100, https://socialmediaarchive.org/record/101, https://socialmediaarchive.org/record/102, https://socialmediaarchive.org/record/103. In addition, we used public data sources for generating map plots and cleaning emojis in hashtags. In particular, for the world map in the main text, the ggplot2 package map$_{d}atafunctionreliesonworldmapdatafrom$. We also downloaded centroid data from https://raw.githubusercontent.com/gavinr/world-countries-centroids/master/dist/countries.csv. For the Congressional district maps in the supplementary repository, we used a Congressional district shapefile from https://www2.census.gov/geo/tiger/GENZ2020/shp/cb_2020_us_cd116_20m.zip and Congressional district code file from https://www2.census.gov/geo/docs/reference/codes/files/national_cd113.txt. To clean emojis in hashtags, we used Unicode data to filer out emojis. The raw dictionary was downloaded from http://www.unicode.org/Public/emoji/1.0//emoji-data.txt. Emoji lists were downloaded from http://www.unicode.org/reports/tr51/index.html#emoji_data. For a comparison to the voting age population, we gathered the voting age population size from https://www.census.gov/content/dam/Census/library/publications/2022/demo/p20-585.pdf.

Code Availability

Code from this study is stored in the Social Media Archive (SOMAR) housed by ICPSR and is accessible under the same terms as the de-identified data. Access to the data can be requested in the following links: https://socialmediaarchive.org/record/87, https://socialmediaarchive.org/record/88, https://socialmediaarchive.org/record/89, https://socialmediaarchive.org/record/90, https://socialmediaarchive.org/record/91, https://socialmediaarchive.org/record/92, https://socialmediaarchive.org/record/93, https://socialmediaarchive.org/record/94, https://socialmediaarchive.org/record/95, https://socialmediaarchive.org/record/96, https://socialmediaarchive.org/record/97,

https://socialmediaarchive.org/record/98, https://socialmediaarchive.org/record/99, https://socialmediaarchive.org/record/100, https://socialmediaarchive.org/record/101, https://socialmediaarchive.org/record/102, https://socialmediaarchive.org/record/103. The data in this study were analyzed using R (version 4.3.3), which was executed on via R notebooks on JupyterLab (4.2.4). The analysis code imports several R packages available on CRAN, including tidyverse (2.0.0), dplyr (1.1.4), ggplot2 (3.5.2), estimatr (1.0.4), Hmisc (5.1-0), sf (1.0-16), ggrepel (0.9.6), patchwork (1.2.0), scales (1.3.0), RColorBrewer (1.1-3), lubridate (1.9.4), stringr (1.5.1) and hbal (1.2.12).

Acknowledgments

The Facebook Open Research and Transparency (FORT) team provided substantial support in executing the overall project. We are grateful for support on various aspects of project management from C. Nayak, S. Zahedi, I. Rosenn, L. Ahmad, A. Bhalla, C. Chan, A. Gruen, B. Hillenbrand, D. Li, P. McLeod, D. Rice, and N. Shah; engineering from Y. Chen, S. Chen, T. Huang, T. Lohman, R. Moodithaya, R. Pyke, L. Saloumi, Y. Wan, and F. Yan; data engineering from S. Chintha, J. Cronin, D. Desai, Y. Kiraly, T. Li, X. Liu, S. Pellakuru, and C. Xie; data science and research from H. Connolly-Sporing, S. Tan, and T. Wynter; academic partnerships from R. Mersey, M. Zoorob, L. Harrison, S. Aisiks, Y. Rubinstein, and C. Qiao; privacy and legal assessment from K. Benzina, F. Fatigato, J. Hassett, S. Iyengar, P. Mohassel, A. Muzaffar, A. Raghunathan and A. Sun; and content design from C. Bernard, J. Breneman, D. Leto and S. Raj. NORC at the University of Chicago partnered with Meta on this project to conduct the fieldwork with the survey participants and pair the survey data with web tracking data for consented participants in predetermined aggregated forms. We are particularly grateful for the partnership of NORC Principal Investigator J.M. Dennis and NORC Project Director M. Montgomery. We thank participants of the Digital Repression Workshop at the Hertie School, participants at the Trust

and Safety Research Conference at Stanford, Stanford Communication Department Colloquium participants, Graphika, DFRLab, Jigsaw, the Social Technologies Lab at Cornell, the Stanford HCI group, the Stanford Social Media Lab and the Stanford Internet Observatory for helpful feedback on earlier versions of this paper.

Meta funded all aspects of project implementation and data collection. Ancillary support (e.g., research assistants and course buyouts), as applicable, was sourced by academics from (authors' initials in parenthesis): the Democracy Fund (N.J.S), the European Research Council Starting Grant (EXPO- 756301) (M.W.), the Guggenheim Foundation (B.Ny.), the Hewlett Foundation (N.J.S, J.A.T.), the Hopewell Fund (J.A.T.), the John S. and James L. Knight Foundation (D.F., D.L., N.J.S, J.A.T., R.T.), the Charles Koch Foundation (J.A.T.), New York University (J.A.T.), the Alfred P. Sloan Foundation (M.G.), Stanford University (H.A., M.G., N.M., J.P., Y.X.), the Stanford Institute for Economic Policy Research (M.G.), the Stanford Freeman Spogli Institute for International Studies (J.P.), The Hoover Institution at Stanford University (Y.X.), the University of Texas at Austin (N.J.S), and the University of Wisconsin–Madison (Y.M.K.). Ruth E. Appel was supported by an SAP Stanford Graduate Fellowship in Science and Engineering, a Stanford Center on Philanthropy and Civil Society PhD Research Fellowship and a Stanford Impact Labs Summer Collaborative Research Fellowship. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

R.E.A., Y.M.K., J.P., Y.X., B.Ni. and D.T. supervised all analyses, and wrote the paper. R.E.A., Y.M.K., J.P., Y.X. were academic lead authors, and Y.M.K., J.P., Y.X. had final control rights. B.Ni. and D.T. were the lead authors at Meta. R.E.A., J.P. and Y.X. analyzed data. R.E.A., Y.M.K., J.P., Y.X., B.Ni. and D.T. designed the study. R.E.A., Y.M.K., J.P., Y.X., D.T., P.B., D.D., D.F., N.M., D.M., B.Ny.,

E.T., R.T., C.V.R., A.W. and M.W. contributed study materials (for example, survey questionnaires, classifiers and software). R.E.A., Y.M.K., J.P., Y.X., B.Ni., D.T., P.B., H.A., D.F., M.G., S.G.-B., A.M.G., S.I., D.L., N.M., B.Ny., J.S., E.T., R.T., M.W., A.C.-T., D.M.,C.V.R., C.K.d.J., A.F., W.M., N.J.S. and J.A.T. contributed to the design of the project. D.T., P.B., T.B., A.C.-T., D.D., D.M., C.V.R., B.X., A.J., A.F. and W.M. coordinated the implementation of the experimental intervention and collected and curated all platform data. B.Ni. coordinated the collection of qualitative deceptive online network descriptions. R.E.A., J.P. and Y.X. contributed the figures and tables. D.L., R.T., M.W., A.C.-T., T.B., B.Ny., P.B., M.G., S.G.-B., A.F., W.M., N.J.S. and J.A.T. provided feedback on the manuscript. N.J.S. and J.A.T. were joint principal investigators for the academic involvement on this project, responsible for management and coordination. C.K.d.J., A.F. and W.M. led Meta's involvement on this project and were responsible for management and coordination.

Competing interests

The costs associated with the research (such as participant fees, recruitment, and data collection) were paid by Meta and some authors are employed by Meta. To ensure transparency and integrity in the research process, we adopted the following conventions. First, none of the academic researchers nor their institutions received financial or any other compensation from Meta for their participation in the project. Second, all of the papers resulting from the U.S. 2020 Facebook and Instagram Election Study, including this one, were preregistered at the Open Science Foundation. Third, for every paper, a set of authors with control rights over the final content of the paper were specified. The authors with control rights for this paper are Young Mie Kim, Jennifer Pan, and Yiqing Xu. Fourth, Meta publicly agreed that there would be no pre-publication approval of papers for publication on the basis of their findings. Finally, we appointed a rapporteur for the project – Professor Michael Wagner of the

University of Wisconsin, Madison – who was neither a paid employee of Meta nor a member of the independent academic research team. For more information, see SI S1.

The following authors are employed by Meta: Ben Nimmo, Daniel Robert Thomas, Pablo Barberá, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Annie Franco, Chad Kiewiet de Jonge, Winter Mason, Devra Moehler, Carlos Velasco Rivera, Arjun Wilkins, Beixian Xiong. Below we list additional declarations from the academic author team: owns Meta Stock (J.S., Y.X.), conducted paid consulting work for Meta (N.J.S.), received direct research funding from Meta (A.M.G., S.I., B.Ny., J.P., J.S., N.J.S., R.T., J.A.T., M.W.), received an honorarium/fee from Meta for attending and/or hosting an event/serving as outside expert (M.G., J.P., J.A.T., N.M.), Attended a Meta event where food, travel, or lodging was paid for by the company (R.E.A., D.F., M.G., S.G.-B., A.M.G., S.I., Y.M.K., D.L., N.M., B.Ny., J.P., J.S., N.J.S., ET, R.T., J.A.T., M.W.), own individual stocks at a related company (JS), received direct research funding from a related company (N.J.S., R.T.), received an honorarium/fee from a related company for attending and/or hosting an event/serving as outside expert (M.G.), attended an event at a related company where food, travel, or lodging was paid for by the company (M.G., D.L., N.M., B.Ny., N.J.S., J.A.T., M.W.). The remaining authors declare no competing interests.

Additional information. Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Jennifer Pan, E-mail: jp1@stanford.edu.

Reprints and permissions information is available at www.nature.com/reprints.

Funding. Meta funded all aspects of project implementation and data collection. Ancillary support (e.g., research assistants and course buyouts), as applicable, was sourced by academics from (authors' initials in parenthesis): the Democracy Fund (N.J.S), the European Research Council Starting Grant (EXPO- 756301) (M.W.), the Guggenheim Foundation (B.Ny.), the Hewlett Foundation (N.J.S, J.A.T.), the

Hopewell Fund (J.A.T.), the John S. and James L. Knight Foundation (D.F., D.L., N.J.S, J.A.T., R.T.), the Charles Koch Foundation (J.A.T.), New York University (J.A.T.), the Alfred P. Sloan Foundation (M.G.), Stanford University (H.A., M.G., N.M., J.P., Y.X.), the Stanford Institute for Economic Policy Research (M.G.), the Stanford Freeman Spogli Institute for International Studies (J.P.), The Hoover Institution at Stanford University (Y.X.), the University of Texas at Austin (N.J.S), and the University of Wisconsin–Madison (Y.M.K.). Ruth E. Appel was supported by an SAP Stanford Graduate Fellowship in Science and Engineering, a Stanford Center on Philanthropy and Civil Society PhD Research Fellowship and a Stanford Impact Labs Summer Collaborative Research Fellowship. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Tables

Figures

Fig. 1 Characteristics of deceptive online networks. Numbers in Panel A reflect the number of networks identified by Meta as originating in each country; transnational networks are shown as individual numbers at the approximate geographic midpoint of their attributed countries of origin and are connected to these countries via dotted lines.

Fig. 2 Reach of deceptive online networks. x-axes are log-transformed due to high concentration (Panels A and B); the sum of the viewers for all networks exceeds the total number of viewers reached on each platform because some viewers were exposed to multiple networks (Panel C); white or grey text labels next to the viewer counts for CIB networks represent the network names in common usage (Panel C).

Fig. 3 How users are exposed to deceptive network content. There are two sources of exposure: network accounts (direct) and non-network accounts (indirect).

Fig. 4 Characteristics of viewers of network content by how users are exposed to network content on Facebook. Horizontal dashed lines correspond to how characteristics are distributed in the population of US adult active users in the study period; Panel D shows number of viewers by source of exposure at the top of each bar. Networks CIB10, CIB11, FMO14, FMO21 and FMO37 are excluded because they had less than 100 viewers in total and no observations in some of the panels.

Fig. 5 User characteristics correlate with exposure/engagement and outcomes: Adding more control variables reduces the correlation between exposure/engagement and outcomes. Top panels refer to FMO27, bottom panels to FMO35. Panels a to f show coefficient estimates and the 95% confidence intervals based on Huber White robust standard errors (HC2) obtained by sequentially adding more pre-exposure covariates to the model and using entropy balancing to adjust for them. Each column shows a different outcome. The number of participants who were exposed to or engaged with deceptive network content are shown for each model in the form of bars on the left of the coefficient plots, the model at the very top also shows exact numbers for orientation. See SI Table S3 for full result details, including p-values and confidence intervals.

References

[1] Stanford Internet Observatory. Reply-Guys Go Hunting: An Investigation into a U.S. Astroturfing Operation on Facebook, Twitter, and Instagram. Tech. rep., Stanford Internet Observatory, https://purl.stanford.edu/vh222ch4142 (2020).

[2] Stanley-Becker, I. Pro-Trump youth group enlists teens in secretive campaign likened to a 'troll farm,' prompting rebuke by Facebook and Twitter. *The Washington Post* Sept. 15 (2020).

[3] Alizadeh, M., Shapiro, J. N., Buntain, C. & Tucker, J. A. Content-based features predict social media influence operations. *Sci. Adv.* **6**, eabb5824 (2020).

[4] Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D. & Volfovsky, A. Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 243–250 (2020).

[5] Ferrara, E., Chang, H., Chen, E., Muric, G. & Patel, J. Characterizing social media manipulation in the 2020 us presidential election. *First Monday* **25**, (2020).

[6] Rid, T. Active Measures: The Secret History of Disinformation and Political Warfare. (Farrar, Straus and Giroux, 2020).

[7] Arif, A., Stewart, L. G. & Starbird, K. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. *Proc. ACM Hum.-Comput. Interact.* **2**, CSCW (2018).

[8] Badawy, A., Ferrara, E. & Lerman, K. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis*

and Mining (ASONAM), 258–265 (2018).

[9] DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J. & Johnson, B. The Tactics & Tropes of the Internet Research Agency. https://digitalcommons.unl.edu/senatedocs/2/ (2019).

[10] Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J. & Tucker, J. A. Exposure to the russian internet research agency foreign influence campaign on twitter in the 2016 us election and its relationship to attitudes and voting behavior. *Nat. Commun.* **14**, 62 (2023).

[11] Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A. & Tucker, J. A. Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. *Int. J. Press/Politics* **25**, 357–389 (2020).

[12] Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J. & François, C. The IRA, social media and political polarization in the United States, 2012-2018. https://digitalcommons.unl.edu/senatedocs/1/ (2019).

[13] Jamieson, K. H. Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know. (Oxford University Press, 2020).

[14] Kim, Y. M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S. Y., Heinrich, R., Baragwanath, R. & Raskutti, G. The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook. *Polit. Commun.* **35**, 515–541 (2018).

[15] Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S. & Tesconi, M. Coordinated Behavior on Social Media in 2019 UK General Election. In *Proceedings of the International AAAI Conference on Web and Social Media*, 443–454 (2021).

[16] Woolley, S. & Howard, P. Computational propaganda worldwide: Executive summary. Tech. rep., Oxford Internet Institute (2017).

[17] Chakraborty, A., Paranjape, B., Kakarla, S. & Ganguly, N. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 9–16 (2016).

[18] Lu, Y. & Pan, J. Capturing Clicks: How the Chinese Government Uses Clickbait to Compete for Visibility. *Polit. Commun.* **38**, 23–54 (2021).

[19] Munger, K. All the News That's Fit to Click: The Economics of Clickbait Media. *Polit. Commun.* **37**, 376–397 (2020).

[20] Cvetkovska, S., Belford, A., Silverman, C. & Feder, J. L. The secret players behind macedonia's fake news sites. *Organized Crime and Corruption Reporting Project* July 18 (2018).

[21] Hughes, H. C. & Waismel-Manor, I. The Macedonian Fake News Industry and the 2016 US Election. *PS: Political Science & Politics* **54**, 19–23 (2021).

[22] Subramanian, S. Inside the Macedonian Fake-News Complex. *WIRED* (2017).

[23] Meta. Inauthentic behavior. https://transparency.meta.com/policies/community-standards/inauthentic-behavior/ (2024).

[24] Gleicher, N. Coordinated inauthentic behavior explained. https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/ (2018).

[25] United States Senate Select Committee on Intelligence. Report on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media with Additional Views. https://www.intelligence.senate.

gov/sites/default/files/documents/Report_Volume2.pdf (2019).

[26] Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539 (2020).

[27] Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. & Watts, D. J. Examining the consumption of radical content on youtube. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101967118 (2021).

[28] Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Malhotra, N., Moehler, D., Pan, J., Thomas, D. R., Tromble, R., Rivera, C. V., Wilkins, A., Wojcieszak, M., Xiong, B., De Jonge, C., Franco, A., Mason, W., Stroud, N. J. & Tucker, J. A. Like-minded sources on facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).

[29] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on twitter during the 2016 us presidential election. *Science* **363**, 374–378 (2019).

[30] Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, M., Xiong, B., De Jonge, C., Franco, A., Mason, W., Stroud, N. J. & Tucker, J. A. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).

[31] Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, M., Xiong, B., De Jonge, C., Franco, A., Mason, W., Stroud, N. J. & Tucker, J. A. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381**, 404–408 (2023).

[32] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E. & Quattrociocchi, W. The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 554–559 (2016).

[33] Flaxman, S., Goel, S. & Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80**, 298–320 (2016).

[34] Hainmueller, J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Polit. Anal.* **20**, 25–46 (2012).

[35] Imbens, G. W. & Rubin, D. B. Causal inference in statistics, social, and biomedical sciences. (Cambridge University Press, 2015).

[36] Imbens, G. & Xu, Y. Lalonde (1986) after nearly four decades: Lessons learned. *arXiv preprint* arXiv:2406.00827 (2024).

[37] Jackson, D. W., Adler, W. T., Dougall, D. & Jain, A. Seismic Shifts: How Economic, Technological, and Political Trends are Challenging Independent Counter-Election-Disinformation Initiatives in the United States. Tech. rep., Center for Democracy & Technology, https://cdt.org/wp-content/uploads/2023/09/seismic-shifts-091823-1302.pdf (2023).

[38] Menn, J., Schaffer, A., Nix, N. & Morse, J. Chinese propaganda accounts found by Meta still flourish on X. *The Washington Post* (2024).

[39] Asmolov, G. The Effects of Participatory Propaganda: From Socialization to Internalization of Conflicts. *J. Des. Sci.* **6**, (2019).

[40] DiResta, R. Invisible Rulers: The People Who Turn Lies into Reality. (PublicAffairs, 2024).

[41] Nemer, D. Disentangling brazil's disinformation insurgency: President jair bolsonaro is already casting doubt on the validity of the 2022 elections. as in 2018, whatsapp is poised to be a key spreader of viral falsehoods. *NACLA Rep. Am.* **53**, 406–413 (2021).

[42] Ong, J. C. & Cabañes, J. V. A. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the philippines. (2018).

[43] Prochaska, S., Duskin, K., Kharazian, Z., Minow, J. I., Blucker, R. T., Venuto, D., West, J. D. & Starbird, K. Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 us presidential election. *Proc. ACM Hum.-Comput. Interact.* **7**, CSCW1 (2023).

[44] Starbird, K., Arif, A. & Wilson, T. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. In *Proceedings of the ACM on Human-Computer Interaction*, CSCW, vol. 3 (Association for Computing Machinery, New York, NY, USA, 2019).

[45] Wanless, A. & Berk, M. Participatory propaganda: The engagement of audiences in the spread of persuasive communications. In Herbert, D. & Fisher-Høyrem, S. (eds.) *Social media and social order*, 111–139 (De Gruyter Open Poland, Warsaw, 2021).

[46] Messing, S. & Westwood, S. J. Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Commun. Res.* **41**, 1042–1063 (2014).

[47] Chaiken, S. & Ledgerwood, A. A theory of heuristic and systematic information processing. *Handbook of theories of social psychology* **1**, 246–266 (2012).

[48] Petty, R. E., Cacioppo, J. T., Petty, R. E. & Cacioppo, J. T. The elaboration likelihood model of persuasion. (Springer, 1986).

[49] Bhadani, S., Yamaya, S., Flammini, A., Menczer, F., Ciampaglia, G. L. & Nyhan, B. Political audience diversity and news reliability in algorithmic ranking. *Nat. Hum. Behav.* **6**, 495–505 (2022).

[50] Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A. & Picard, R. Human detection of political speech deepfakes across transcripts, audio, and video. *Nat. Commun.* **15**, 7629 (2024).

[51] Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R. & Hertwig, R. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nat. Hum. Behav.* **4**, 1102–1109 (2020).

[52] Facebook. October 2020 Coordinated Inauthentic Behavior Report. Tech. Rep. October, Facebook, https://about.fb.com/wp-content/uploads/2020/11/October-2020-CIB-Report.pdf (2020).