Black trolls matter

Racial and ideological asymmetries in social media disinformation

Freelon, Deen; Bossetta, Michael; Wells, Chris; Lukito, Josephine; Xia, Yiping; Adams, Kirsten

Published in:
Social Science Computer Review

DOI:
10.1177/0894439320914853

2022

Document Version:
Peer reviewed version (aka post-print)

Black trolls matter: Racial and ideological asymmetries in social media disinformation

(accepted for publication in *Social Science Computer Review*, doi: 10.1177/0894439320914853)

Deen Freelon[a]*, Michael Bossetta[b], Chris Wells[c], Josephine Lukito[d], Yiping Xia[d], and Kirsten Adams[a]

[a]School of Media and Journalism, University of North Carolina, Chapel Hill, USA; [b]Department of Political Science, University of Copenhagen, Copenhagen, Denmark; [c]College of Communication, Boston University, USA; [d]School of Journalism and Mass Communication, University of Wisconsin, Madison, USA

*Corresponding author:

Deen Freelon
Hussman School of Journalism and Media
University of North Carolina at Chapel Hill
Carroll Hall, CB 3365
Chapel Hill, NC 27599
Freelon@email.unc.edu
@dfreelon

This article has not been previously rejected by another journal.

This article is at least 80% different from other articles we have published, and no table or figure has appeared in any other article.

Abstract: The recent rise of disinformation and propaganda on social media has attracted strong interest from social scientists. Research on the topic has repeatedly observed ideological asymmetries in disinformation content and reception, wherein conservatives are more likely to view, redistribute, and believe such content. However, preliminary evidence has suggested that race may also play a substantial role in determining the targeting and consumption of disinformation content. Such racial asymmetries may exist alongside, or even instead of, ideological ones. Our computational analysis of 5.2 million tweets by the Russian government-funded "troll farm" known as the Internet Research Agency sheds light on these possibilities. We find stark differences in the numbers of unique accounts and tweets originating from ostensibly liberal, conservative, and Black left-leaning individuals. But diverging from prior empirical accounts, we find racial presentation—specifically, presenting as a Black activist—to be the most effective predictor of disinformation engagement by far. Importantly, these results could only be detected once we disaggregated Black-presenting accounts from non-Black liberal accounts. In addition to its contributions to the study of ideological asymmetry in disinformation content and reception, this study also underscores the general relevance of race to disinformation studies.

Keywords: disinformation, Internet Research Agency, Twitter, ideological asymmetry, digital blackface

Since the 2016 US presidential election, political disinformation has taken center stage in social media research (Freelon & Wells, in press). A flurry of studies has emerged since then to explain the what, how, and why of disinformation on social media (e.g. Benkler, Faris, & Roberts, 2018; Faris et al., 2017; Howard, Ganesh, Liotsiou, Kelly, & Francois, 2018; Jamieson, 2018). The importance of this phenomenon stems from two foundational, though rarely stated, norms undergirding American political communication. First, the content of political messages should be factual to the extent that independent observers can agree on the facts. As an example that violates this norm, flyers that spread false information about voting times and procedures have been a problem for decades (Daniels, 2009). Second, political communicators should represent their identities and intentions honestly. Those who fail to do so imply that deception is a key element of their communication strategy, and that without it, their messages would lose their efficacy.

One of the most prominent theoretical perspective applied in studies of disinformation, misinformation, "fake news" and related phenomena is *ideological asymmetry*. This idea posits that individuals occupying the left and right poles of an assumed unidimensional axis of ideology are targeted by, engage with, and/or believe such content disproportionately. Existing research has repeatedly identified the American right wing as substantially more vulnerable to disinformation attack and acceptance than the left (Allcott & Gentzkow, 2017; Benkler et al., 2018; Grinberg et al., 2019; Guess et al., 2019). Extending this line of inquiry, we investigate the possibility of both ideological and racial asymmetry in a disinformation campaign executed before, during, and after the 2016 US elections by company working on behalf of the Russian government. Prior research has provided preliminary evidence of such a racial asymmetry but

did not investigate the question directly (DiResta et al., 2018; Howard et al., 2018). The current study does so and detects racial and ideological asymmetries in both disinformation content and reception.

The organization responsible for the digital disinformation campaign analyzed here is known as the Internet Research Agency (IRA), a so-called "troll farm" funded by the Russian government and based in St. Petersburg. While much of the emerging academic research on disinformation has focused on perpetrators' impersonation of news outlets and content (Allcott & Gentzkow, 2017; Grinberg et al., 2019; Guess et al., 2019; Vargo et al., 2017), one of the IRA's key strategies is to create social media accounts called "sockpuppets" that purport to be American citizens with strong political interests. Among other identity categories, IRA agents masqueraded as liberal and conservative Twitter users, but targeted left-leaning Black users separately from the former (DiResta et al., 2018; Howard et al., 2018). In so doing, they engaged in a practice known as "digital blackface" (Green, 2006; Robertson et al., 2018; Stark, 2018), in possibly its first use as a disinformation tactic. Thus, this study connects the disparate literatures of racial impersonation and digital disinformation by demonstrating that, at least in the American context, race is a key vulnerability ripe for exploitation by disinformation purveyors.

Asymmetries in disinformation content and reception

Before we begin reviewing the literature on digital disinformation and propaganda, some definitional clarification is in order. Imprecise scholarly and journalistic uses of terms such as "fake news," "misinformation," "disinformation," "propaganda," "junk news," etc. have spread confusion about the messages and behaviors to which those terms are attached. Therefore, this study follows the recommendation of the European Commission's High Level Expert Group on

Fake News and Online Disinformation on the appropriate terminology for deceptive online content: "…we favour the word 'disinformation' over 'fake news.' Disinformation… includes all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit" (2018, p. 3). This definition covers a wide variety of content that, unlike traditional political communications, has essentially been weaponized to inflict harm by any means available.

As Freelon and Wells (in press) document, the study of disinformation is still in its infancy, having seen only infrequent and scattered scholarly attention prior to 2017. The explosion of studies on the topic following the 2016 US elections coalesced quickly around the theory of *ideological asymmetry*. Studies that examine ideological asymmetry in disinformation contexts build on a rich research tradition that has examined the concept's application to a range of psychological outcomes (e.g. Collins et al., 2017; Crawford, 2014; Grossmann & Hopkins, 2015; Jost et al., 2017). Technically the term can be used to describe any behavioral or cognitive difference between individuals holding opposed ideologies; but in the disinformation literature, it usually refers to a disproportionate tendency among conservative individuals to encounter, redistribute, and/or believe disinformation. Here we distinguish between two distinct dimensions of ideological asymmetry with respect to disinformation: the *content* dimension, wherein disinformation purveyors produce vastly more conservative-oriented than liberal messages; and the *reception* dimension, in which conservatives who view disinformation targeted at them are more likely to engage with it. Empirical studies have found ample support both for ideological asymmetry's content dimension (Allcott & Gentzkow, 2017; Benkler et al., 2018; DiResta et al., 2018; Howard et al., 2018) and its reception dimension (Badawy et al., 2018; Freelon & Lokot, 2020; Grinberg et al., 2019; Guess et al., 2019; Hjorth & Adler-Nissen, 2019). In other words,

disinformation agents have been empirically found to produce more content targeting conservatives than liberals, and conservatives are also more likely to engage with the disinformation content targeted at them than are liberals.

The existence of disparities between liberals and conservatives in disinformation content and reception raises the question of what other identities may suffer disproportionate risk of disinformation targeting and engagement. The scholarly interest in ideological asymmetry may be considered one case within a more general project of understanding the different disinformation risk profiles of distinct social groups. Based on existing studies of the IRA specifically, we suggest that disinformation content and/or reception may also accrete unequally along ethnic and racial lines. In particular, we argue that such racial asymmetries cannot be reduced to those based on ideology. Our data afford us a unique opportunity to test the respective magnitudes of one key asymmetry—that between Blacks and Whites—to that between ideological opponents on the left and right.

Digital blackface and racial asymmetry

To properly conceptualize both the ideological and racial asymmetries that may surround the IRA's activities, it is important to discuss *sockpuppetry,* which is the practice of creating false social media identities in order to conceal one's true identity and motives. The IRA's primary method of attack was to create sockpuppet accounts posing as members of various American social, political, and racial groups. Such practices are distinct from parody and satirical accounts, which often impersonate celebrities but are clearly labeled as inauthentic.[1] Sockpuppetry is a type of trolling (Buckels et al., 2014) in that its goal is to annoy, harass, or otherwise harm its targets. The measure of a sockpuppet's success is the extent to which

audiences are fooled into believing the false identity is the real one. A related term that is applied mostly to attempt to create false grassroots organizations is *astroturfing*, which occurs both on and off social media (Peng et al., 2017; Ratkiewicz et al., 2011). Most of the currently available information on the IRA's sockpuppet identities is descriptive, with both Howard et al. (2018) and DiResta et al. (2018) listing examples of conservative, progressive, African-American, and Muslim-American identities. Linvill and Warren (2018) usefully place most of the known IRA accounts into one of eight categories, and this study's empirical analysis builds on their work. The major difference between Linvill and Warren's typology and ours is that we separate Black (African American)-presenting accounts into their own category, whereas they included such accounts with other "Left Trolls." Of the many social identities the IRA impersonated, we focus on Black people for several reasons. First, along with conservatives and progressives which are already represented in Linvill and Warren's typology, Black people are repeatedly discussed in the IRA research literature as targets of substantial propaganda efforts. DiResta et al. write that "the IRA created an expansive cross-platform media mirage targeting the Black community" and "the degree of integration into authentic Black community media was not replicated in the otherwise Right-leaning or otherwise Left-leaning content" (2018, p. 8). Howard et al. find that on Twitter, "the IRA focused their political messaging on two targets above others: conservative voters and African Americans" (2018, p. 26). The Mueller report cites multiple examples of the IRA's attempts to target Black Americans through social media (Mueller, 2019, pp. 22, 25). These preliminary reports raise the possibility that a racial asymmetry between the IRA's Black and White-presenting accounts may exist alongside or possibly instead of the well-documented ideological asymmetries found in other disinformation campaigns.

The second reason to analyze Black impersonation for disinformation purposes is to contribute to the small but growing literature on digital blackface. This term refers to digitally-mediated renderings of online selves as Black by non-Black individuals (Dobson & Knezevic, 2018; Green, 2006; Robertson et al., 2018; Stark, 2018). Like its 19th-century counterpart, digital blackface appropriates Black visages for the self-expression or entertainment of non-Black audiences. Existing research focuses on such examples as photographic filters that make subjects appear "Blacker" (Stark, 2018), image macros and GIFs featuring Black individuals (Dobson & Knezevic, 2018), dark-complexioned emojis (Robertson et al., 2018), and Black video game avatars (Green, 2006). These studies tend to address the cultural politics of digital blackface, typically focusing on how it operates as a manifestation of anti-Black racism regardless of the perpetrator's intentions.

In the current context, digital blackface operates as a type of sockpuppetry that specifically impersonates Black identities. By combining the deceptive intentionality of the former and the distinctive visual and linguistic markers of the latter, the IRA fashioned a malicious technique to exploit American racial divisions for geopolitical advantage. Given the novelty of this phenomenon, most of what we currently know about it comes from news reports and the foundational (but non-peer-reviewed) research of DiResta et al. (2018) and Howard et al. (2018). Aside from the IRA, sockpuppets in digital blackface have also been documented as working on behalf of men's rights activists (Broderick, 2014) and unidentified pro-Trump parties (Weill, 2019). The current study is the first to analyze digital blackface at scale in search of Black/White racial asymmetries in disinformation content and reception.

Our first three research questions concern the relative sizes of such racial asymmetries compared to those based on ideology.

- RQ1: Do ideological and/or racial asymmetries exist in the numbers of IRA Twitter accounts posing as conservatives, liberals, and Black users?
- RQ2: Do ideological and/or racial asymmetries exist in the numbers of messages produced by IRA Twitter accounts posing as conservatives, liberals, and Black users?
- RQ3: Do ideological and/or racial asymmetries exist in audience engagement (retweets, likes, and replies) with IRA Twitter accounts posing as conservatives, liberals, and Black users?

False amplification, false asymmetries?

Before proceeding to the empirical sections of our study, we need to address a theoretical possibility that, if true, would completely reframe the relevance of our findings. Weedon, Nuland, and Stamos (2017) define *false amplification* as "coordinated activity by inauthentic accounts with the intent of manipulating political discussion" (p. 6). While this definition could arguably include all the IRA's social media activities, we are especially concerned with the prospect that its agents could generate false reception patterns by retweeting, favoriting, and/or replying to its own content. If apparent asymmetries along ideological, racial, or other dimensions can be explained by this type of false amplification, they cannot be called true asymmetries at all. Rather, they would represent what we might call *false asymmetries*, possibly intended to make sockpuppet accounts look more popular and influential than they are in the eyes of authentic users.

The practice of purchasing perceived influence indicators such as followers, likes, retweets, etc. has long been studied by computer science and information scientists (e.g. Aggarwal & Kumaraguru, 2015; Singh et al., 2016; Stringhini et al., 2013). Users who patronize

such services are usually seeking instant popularity—an attractive prospect given the low price of high volumes of false digital interactions. Research into politically motivated false amplification is still in its infancy, and very few studies have empirically addressed the phenomenon. A few studies have examined political botnets, or networks of automated social media accounts mobilized in support of a political goal. Examining a Twitter botnet focused on Brexit in 2016, Bastos and Mercea (2019) find that its component accounts spent much of their time retweeting hyperpartisan content originating from both authentic users and other bots. Hegelich and Janetzko (2016) find similar results for a Twitter botnet tweeting about the Ukraine/Russia conflict in early 2014. In documenting a Venezuelan botnet active in 2015, Forelle et al. (2015) conclude that bots are minor players in the national Twitter conversation and mostly retweet content originally posted by politicians.

Because the IRA used manually-controlled accounts alongside automated ones (Mueller, 2019, pp. 27–29), it is not currently known how closely it will follow the botnet-style false amplification playbook. Our final research question probes this possibility, and in so doing offers evidence about the extent to which any asymmetries we may find are produced by authentic users.

- RQ4: To what extent do IRA Twitter accounts engage in false amplification by retweeting, favoriting, and replying to their own messages?

Data and methods

Background

The Internet Research Agency represents a highly apt case for evaluating the present research questions for several reasons. First, as we detail below, more data are available about them than any other English-language social media-based disinformation campaign. Such data

are difficult to obtain because social media platforms typically remove the content immediately upon detection. Second, the scope of the IRA's campaign enables comparisons between different tactics: rather than focusing on a single issue, their targets and topics covered a broad swath of the American political landscape. Third, available evidence indicates that at minimum, tens of millions of Americans may have viewed IRA content (DiResta et al., 2018; Howard et al., 2018; Mueller, 2019). The scale of their reach alone merits a thorough investigation of their activities. Fourth, our results will offer a baseline against which future disinformation campaigns may be compared, so that we can better understand which characteristics are specific to the IRA and which generalize beyond them.

The IRA's sockpuppet tactics placed them into direct conflict with Twitter's terms of service, which forbid the use of "misleading account information in order to engage in spamming, abusive, or disruptive behavior, including attempts to manipulate the conversations on Twitter" (Twitter, n.d.). Consequently, Twitter summarily suspended IRA accounts as it detected them, in the process eliminating researcher access. On November 1, 2017, the US House Intelligence Committee posted a PDF containing the 2,752 known (at the time) screen names of suspected IRA accounts that Twitter had provided to it. This information enabled certain types of research, such as analyses of IRA activity within datasets researchers had collected prior to the mass account suspension (e.g. Badawy et al., 2018; Broniatowski et al., 2018; Hjorth & Adler-Nissen, 2019). Unfortunately, such serendipitous access did not permit a broad analysis of the IRA's tactics across topics and account types. This did not become possible until Twitter, in an unprecedented move, began publicly releasing disinformation-related datasets in late 2018.

Dataset

On October 17, 2018, Twitter announced two datasets containing tweets posted by suspected IRA accounts. It posted one dataset publicly, but redacted the screen names, display names, and user IDs of all accounts that had accrued fewer than 5,000 followers at the time of account suspension. Of the 3,667 unique screen names present in this dataset, only 167 (4.6%) were left unredacted. This dataset was not suitable for the current study, because we needed to know each author's screen name and display name to categorize each tweet according to its author's sockpuppet identity. Fortunately, Twitter also made an unredacted version of the dataset available to researchers who fill out an application detailing how they plan on using it. We submitted this application and were granted access to Twitter's official, unredacted IRA dataset on March 29, 2019.

This dataset contains 8,768,633 tweets posted by 3,479 unique users between May 9, 2009 and June 21, 2018.[2] It is by far the largest IRA dataset available, surpassing earlier public datasets posted by NBC News (Popken, 2018) and Linvill and Warren (2018). Additionally, because the official Twitter dataset was compiled after the IRA accounts had been suspended, its retweet, like, and reply counts can be considered definitive. In contrast, the other two datasets were collected soon after each tweet was posted, so they would have omitted any reactions accrued after that point.

[Table 1 here]

Identity categories

To categorize the IRA users, we adapt Linvill and Warren's (2018) eight IRA user categories, which generally align with both published journalistic accounts and our own unpublished category scheme (Table 1).[3] The eight categories collectively include 3,210 unique

users, including those filed under the catchall categories of "Non-English" and "Unknown." After applying their categories to the screen names in the dataset, we found that they cover 85.3% of the unique users and 97.2% of the tweets.

The major limitation of Linvill and Warren's categories is that they do not separate Black-presenting users from generic left-identifying users. To do so, two of the authors and a third coder manually and independently re-analyzed the screen names, descriptions, and a random selection of five tweets from all users in the Left Troll category to determine which showed evidence of digital blackface.[4] Using the criterion that two coders had to agree for affirmative judgments to hold, we found 107 of the original 218 Left Trolls (49%) to be Black-presenting. We then moved these users from the Left Troll category into a newly-created Black Troll category, resulting in a total of nine categories.

Prior to analysis, we generated several additional covariates using custom-written R code, including the numbers of hashtags, URLs, and screen names mentioned per tweet; tweet age in days; each account's age; the numbers of followers and followed per account; tweet length in characters, and the presence of images and videos (including animated GIFs). We also removed all retweets initiated by IRA accounts so as to model only the retweets, reply, and like counts attached to tweets originally posed by the IRA. The resulting dataset contains 5,202,233 tweets, all categorized according to both Linvill and Warren's original scheme and our modified version with the Black Troll category. This represents a majority (59.3%) of the original uncategorized dataset, with the largest share of the proportional reduction by far (37.9% of the original dataset) accounted for by IRA-initiated retweets.

Results

[Figure 1 here]

To answer RQ1 and RQ2, we tallied the numbers of users and tweets accounted for by each user category. Figures 1 and 2 show that a substantial proportion of the IRA's efforts were directed at non-English language audiences: 40.5% of unique users posted primarily in languages other than English, with Russian being most popular by far; and a majority of tweets (59.2%) were written in non-English languages. Turning to accounts tweeting in English (and ignoring the Unknown category, about which little can be said), Right Trolls clearly outnumber Left Trolls and Black Trolls combined, with the latter two categories being nearly equal in size (111 and 107 accounts respectively).

However, because accounts may tweet at different rates, an analysis of tweet volumes (RQ2) could tell a somewhat different story. This turns out to be the case: while Right Troll tweets outnumber Left- and Black Troll tweets combined, Black Trolls tweeted nearly three times more than Left Trolls and slightly over a third as often as Right Trolls. Viewed in isolation, this looks like evidence of both racial and ideological asymmetry on the content side, but these results should be interpreted in context with the answers to the remaining research questions. We note in passing that the second smallest category in terms of unique accounts, NewsFeed, was also the second most prolific, further underscoring the lack of correlation between the two metrics.

[Figure 2 here]

To address the possibility of racial and/or ideological asymmetry within the reception dimension (RQ3), we analyze the numbers of retweets, likes, and replies each account category attracted. Over the more than nine years this study covers, IRA accounts were liked a total of 36,064,169 times, retweeted 31,107,778 times, and replied to 2,524,657 times. As with tweet production, these reaction types were distributed unevenly across the account categories. Our

first analysis of the reaction data is descriptive and intended to give an overview of this distribution. Because a large majority of IRA tweets received no reactions (79.8% for retweets, 84.7% for likes, and 92.6% for replies), typical indicators of central tendency like means or medians would be misleading. Instead, Figure 3 shows the proportions of each category's tweets that received at least one of each reaction type. For example, 75.1% of Black Troll tweets attracted at least one like, 73.1% attracted at least one retweet, and 42.9% attracted at least one reply. This puts the Black Troll category in the top position for all three reaction types, with Right Troll a distant second in each case. Black Troll was the only category to elicit at least one like and retweet for more than 50% of its tweets. Of the other categories, Right Troll, Hashtag Gamer, Left Troll, and Newsfeed generally manage to attract at least one like and retweet for between 35% and 20% of tweets, and at least one reply for between about 25% and 5% of tweets. The remaining categories received only minuscule amounts of attention, with Fearmonger accounts receiving zero likes for 98.5% of their tweets and even smaller proportions of retweets and replies.

[Figure 3 here]

Conducting a more rigorous test of which account types were associated with higher and lower reaction activity required an unorthodox statistical approach. The negative binomial regression models we applied to the full dataset all failed to converge due to its large size and extremely long-tailed distributions in the outcome variables. To overcome this methodological obstacle, we employed a repeated subsampling procedure consisting of the following steps:

1. For the dataset in which Black Trolls were separated from Left Trolls, draw a stratified random sample of 500 tweets each from all nine account types, resulting in a total sample size of 4,500.
2. Do the same for the dataset in which Black Trolls are combined with Left Trolls.
3. Compute negative binomial models predicting retweet, like, and reply counts for each of the two samples (six models total: three response metrics by two model sets each), using the "Fearmonger" type as the dummy reference category.[5]
4. Repeat steps 1-3 1,000 times.
5. Compute means and confidence intervals for the negative binomial regression coefficients, standard errors, and p-values within each of the six 1,000-model sets.

The resulting model sets offer stronger evidence than a single sample would, as the confidence intervals for the coefficients indicate the range over which the latter varied. Conveniently, the coefficients, p-values, and standard errors were all normally distributed, making the mean an appropriate indicator of central tendency.

[Table 2 here]

Tables 2, 3, and 4 display the model sets' output, with each table combining both model sets that predict a single response metric. We begin by considering the model set 1s for the three metrics, in which Black Trolls are separated from Left Trolls. We measure ideological asymmetry by subtracting each Left Trolls coefficient from that of Right Trolls, and racial asymmetry by subtracting the Black Trolls coefficient from Left Trolls' (since the latter two are theoretically likely to be closer due to their ideological proximity). Under these definitions, racial asymmetry is evident within the model set 1s for all three metrics, while ideological asymmetry

appears in two (retweets and likes—the Left Troll and Right Troll confidence intervals overlap for replies). Specifically, we observe that the racial asymmetry for retweets is 2.30 times larger than its ideological asymmetry, and that this ratio is 0.91 for likes and 3.29 for replies. Thus, for two of the three metrics (retweets and replies), the racial asymmetry is well over twice as large as the ideological asymmetry, and for the third (likes) it is nearly the same size.

Comparing each metric's model set 2 to its corresponding set 1 demonstrates the importance of separating Black Trolls from Left Trolls. In each case, doing so drastically diminishes the size of the ideological asymmetry, in one case eliminating it entirely. The consistent story here is that much of the Left Troll category's ability to attract attention depends upon Black Trolls being combined with it. We can quantify the impact of this categorical aggregation by dividing the coefficient differences between Left Trolls and Right Trolls in each model set 1 by the corresponding differences in each model set 2, subtracting the resulting quotients from one, and multiplying by 100 to obtain percentages. Thus, for retweets, we conclude that 64.9% of the ideological asymmetry in model set 2 can be attributed to the effect of Black Trolls. The corresponding values for likes and replies are 42.3% and 69.6% respectively.

[Table 3 here]

The remaining account categories present a less consistent story. Hashtag Gamer, Commercial, and Non-English tweets display a mix of significant and non-significant results across the model sets and coefficients, with the magnitude of each effect varying. The NewsFeed category is not significant in any of the six model sets, indicating that its ability to garner attention was indistinguishable from that of the false news-spreading Fearmonger accounts. The presence of image and video content tends to attract engagement, while URL count is negatively

associated with it for two of the three metrics. Most of other the control variables show non-significant or very small effect sizes.

[Table 4 here]

The relevance of the results reported above depend upon most of the response metrics emanating from real people. If the IRA devoted substantial resources to retweeting, liking, and replying to one another's posts—in other words, to false amplification—the above analysis would not be valid. We tested this possibility (RQ4) by investigating the proportions of retweets and replies *to* IRA tweets contributed *by* IRA accounts of all retweets and replies received.[6] To do so, we took advantage of the *retweet_tweetid* and *in_reply_to_tweetid* fields, which show the original tweet IDs to which every IRA retweet and reply responded, respectively. We counted every time one IRA account retweeted or replied to another by cross-indexing these fields against the *tweetid* field. We then divided the respective counts of IRA-to-IRA retweets and replies by the sums of all retweets and replies received by all IRA accounts.

Our analysis detected minimal amounts of false amplification for retweets and replies. Approximately 2.89% of all retweets received and 1.27% of replies received came from other IRA accounts. These findings are consistent with the conclusion that most of the attention the IRA's Twitter accounts received came from authentic users rather than the IRA itself.

[Table 5 here]

Even if we accept this result as true, some account categories might be falsely amplified to disproportionate degrees. To account for this possibility, we extended the above analysis to search for false amplification within each account type—that is, the rates of retweets/replies received by each account type from all other IRA accounts of all within-type retweets/replies (Table 5). As it happens, certain account categories receive large majorities of their responses

from IRA accounts, while others receive almost none from them. For example, Black-, Left-, and Right Trolls each received well under 1% of their retweets and replies from the IRA, but Fearmongers got almost three-quarters of their retweets from them, while Unknown accounts received 63%. There was generally more coordination with retweets than with replies, as the latter maximum was only 11% (for NewsFeeds).

Discussion

We open this section by exploring the implications of the answers to our four research questions.

RQ1 and RQ2

We consider the answers to RQ1 and RQ2 together as their theoretical implications are similar. Ideological asymmetry is readily apparent in the numbers of both unique accounts and tweets by those accounts. To begin with the former, there were well over twice as many Right Trolls as there are Left Trolls and Black Trolls combined. Specifically, Left Troll plus Black Troll accounts amount to just over half the total number of Right Troll accounts (54.0%). This is especially striking in light of recent studies finding that liberals outnumber conservatives on Twitter (Freelon, 2019; Wojcik & Hughes, 2019). It also qualifies as a racial asymmetry given that there were far fewer Black Troll than Right Troll accounts, and close to as many of the former as Left Troll accounts. We find similar evidence of ideological and racial asymmetry in the volume of tweets produced by the three political types of accounts. The numbers of Left Troll plus Black Troll tweets amount to 50.6% of Right Troll tweets, a result similar in magnitude to that for unique accounts. Again, we see that the ideological asymmetry skews toward conservative-targeting content. Interestingly, the number of tweets generated by Black

Trolls exceeds those of Left Trolls but falls below Right Troll content, which should be interpreted in light of the substantial overrepresentation of Black users on Twitter (Perrin & Anderson, 2019).

Overall, these findings dovetail with prior research that has detected ideological asymmetry in the production of disinformation content (Allcott & Gentzkow, 2017; Badawy et al., 2018; Benkler et al., 2018; DiResta et al., 2018; Howard et al., 2018). But it also broadens this nascent literature by demonstrating that racial asymmetries in content production can exist alongside ideological ones. The rightward skew of the ideological asymmetry is consistent with the inference that the IRA considers the American right wing to be especially vulnerable to disinformation attacks. Research on the psychological antecedents of disinformation susceptibility has also found that it skews ideologically rightward (Benkler et al., 2018, p. 328; De keersmaecker & Roets, 2019; Guess et al., 2019; Jost et al., 2018). However, our findings on the consumption side contradict these prior studies, as we explain below.

RQ3

In terms of consumption, we find that despite there being substantially more conservative-presenting accounts and tweets than liberal- or Black-presenting ones, when considered on a per-tweet basis, Black-presenting IRA accounts attract more retweets, likes, and replies than any other identity category. We further discover that the magnitude of the ideological asymmetry between conservative and liberal accounts decreases massively when Black Trolls are separated from the latter. The model sets that separate Black Trolls from Left Trolls reveal that the engagement coefficients of the latter are much closer to those of Right Trolls than those that combine the former two. This indicates that some of the social media engagement dynamics that

appeared to earlier researchers as ideological asymmetry (e.g. in Badawy et al., 2018; Benkler et al., 2018) may have been driven to a significant extent by race. The fact that this analysis was conducted on the highest-quality IRA dataset currently available lends further credence to our conclusion. Moreover, our results reveal that conservative-presenting accounts attracted less engagement per tweet than either liberal- or Black-presenting accounts. Conservative tweets did garner the most engagement in total (25,040,986 combined retweets, likes, and replies; compared to 22,704,973 for Black Trolls and 4,577,743 for Left Trolls), but that was due to the disproportionate size of the IRA's conservative-targeting operation relative to its other operations. Additional research will be needed to determine the extent to which this effect is specific to the IRA, Twitter, or both.

Another major implication of this cluster of findings is that race is a critical variable in the analysis of disinformation uptake and should be a key focus area in future research on the topic. Early inductive research on the IRA first revealed race as a notable determinant of IRA strategy, and this study supports the notion that authentic users interacted with Black-presenting IRA accounts disproportionately relative to other types of accounts. Had we relied on prior typologies of IRA accounts, we would have been unable to detect this result. We should note at this point that Russian exploitation of American racial division is not new—rather, it is the latest iteration of a nearly century-old tradition (Roman, 2019). But whether it is Khrushchev exploiting Jim Crow to claim the moral high ground over the US in the 1950s or IRA agents engaging in digital blackface on Twitter, the goal of undermining American democracy remains constant. Those who seek to build on this study are advised to investigate whether other types of disinformation campaigns draw on the IRA playbook or otherwise exploit ethnoracial divisions, Black/White or otherwise.

These results also link the formerly disparate literatures on digital blackface and state-sponsored social media disinformation. Whereas the former was discussed primarily as an affront to Black cultural sensibilities, we reveal it here as a highly effective tactic in an international disinformation campaign. Whether digital blackface will continue to serve such a function in future political contexts or emerge in other disinformation agents' repertoires (state-sponsored or otherwise) remains to be seen. Either way, its use in the current context should concern not only members of the impersonated group, but also everyone who wants to ensure that American democracy remains free of surreptitious foreign influence.

Before proceeding to our final research question, we want to clarify exactly what we claim to have measured here and how it differs from previous research. Our statistical models reveal substantial racial and ideological asymmetries in online engagement with social media disinformation content. While this activity occurs squarely within the reception dimension and reveals much about the tactical efficacy of racial and ideological sockpuppetry, it does not tell us anything directly about the ideological or racial identities of the individuals who engaged with the accounts (cf. Badawy et al., 2018; Hjorth & Adler-Nissen, 2019). In other words, knowing that tweets by Black-presenting IRA accounts accrued more retweets, likes, and shares than any other category does not in itself justify the claim that Black people were more likely than non-Blacks to engage with IRA accounts. However, such a claim would be consistent with the preexisting findings that 1) most people who replied to IRA accounts shared the ideological dispositions those accounts falsely exhibited (Freelon & Lokot, 2020) and 2) the IRA devoted a disproportionate amount of its Facebook advertising budget to microtargeting Black users (Howard et al., 2018). All this evidence taken together is highly suggestive but far from

conclusive on the question of potential racial asymmetries in disinformation audiences; therefore, future studies should address it directly.

RQ4

Our search for false amplification revealed little evidence thereof. IRA accounts generated very small proportions of the retweets and replies they received, implying that most such responses came from non-IRA accounts. While this finding alone does not prove that most IRA retweets and replies came from authentic users, it does eliminate one major objection to that conclusion. In answering this research question, we identified a specific subtype of false amplification wherein sockpuppet accounts redistribute their own content to increase its prominence. The fact that this occurred so infrequently suggests that the IRA is more interested in convincing real users to spread their messages. Researchers interested in mapping disinformation propagation tactics should consider searching for evidence of such internal content redistribution within other campaigns. To that end, the method we use to identify this activity can be applied to other datasets that contain the required metadata fields. Other research has explored the phenomenon of inauthentic accounts retweeting authentic users (Bastos & Mercea, 2019; Forelle et al., 2015), which is a different kind of false amplification that also deserves further study.

Limitations

Despite the above contributions, our study's limitations should be acknowledged. First, the only identity category we reclassified from Linvill and Warren's typology accounted for sockpuppets that impersonated Black American users. While we stand by this decision based on the resources the IRA devoted to Black sockpuppetry, other sockpuppet identities were present as well,

including Muslim Americans, LGBTQ Americans, and the Latinx community. Whether the IRA fabricated enough such accounts to support a quantitative analysis is an empirical question for future research. Second, because our analysis is cross-sectional, we cannot draw any conclusions about the extent to which the different account types' levels of visibility may have changed over time. Howard et al. (2018, p. 9) first observed substantial English-language IRA activity in 2014, but beyond that, we know little about the specific events or circumstances that allowed some account types to attract relatively large audiences. Third, our methods offer little insight into the IRA's internal strategies and decision-making processes. The best-performing account types were also among the least prolific, which raises the question of whether the IRA learned from its own experiences about which tactics worked best and incorporated those lessons into future actions. Answering questions such as these might require qualitative interviews with ex-IRA employees or others with knowledge of the organization's internal operations.

This study points to a number of opportunities for future research to build on its findings. Aside from those already mentioned, it would be worthwhile to explore how authentic users responded to the IRA attack: for example, did they exhibit evidence of polarization, false beliefs, or diminished confidence in political institutions? And were these effects ideologically or racially asymmetrical? Researchers interested in such questions can take advantage of the fact that although the IRA's tweets have all been removed from Twitter, many of the replies to those tweets are still present (see Freelon & Lokot, 2020). Further examination of racial and ideological asymmetries in disinformation on other social media platforms would help determine which are and are not platform-specific. Such research would also build on DiResta et al. (2018) and Howard et al. (2018), but unfortunately, much of their non-Twitter social media data are not publicly available. Finally, perhaps the most pressing question for future research addresses the

extent to which our findings generalize beyond the case of the IRA. Twitter has publicly posted datasets of non-IRA disinformation campaigns, so it would be interesting to explore, for example, whether these other campaigns also exploited local ideological and/or ethnoracial divisions to spread their content. Many of the tweets in these datasets are written in non-English languages, so linguistic and cultural expertise will be essential in conducting such research.

Author information

Deen Freelon, freelon@email.unc.edu

Michael Bossetta, mjb@ifs.ku.dk

Chris Wells, cfwells@bu.edu

Josephine Lukito, jlukito@wisc.edu

Yiping Xia, xia37@wisc.edu

Kirsten Adams, kkirsten@live.unc.edu

Data availability

The data analyzed in this study may be obtained by submitting a written request to Twitter via the following URL: https://transparency.twitter.com/en/information-operations.html

Software information

The lead author wrote several R scripts to perform the analyses described in this study. They can be downloaded here: [link to be created upon article acceptance]

References

Aggarwal, A., & Kumaraguru, P. (2015). What they do in shadows: Twitter underground follower market. *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, 93–100. https://doi.org/10.1109/PST.2015.7232959

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265.

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38–54. https://doi.org/10.1177/0894439317734157

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Bonn, T. (2019, October 7). *Poll: Overwhelming majority of black voters back any 2020 Democrat over Trump* [Text]. TheHill. https://thehill.com/hilltv/rising/464680-poll-overwhelming-majority-of-black-voters-choose-any-given-2020-democrat-over

Broderick, R. (2014, June 17). *Activists Are Outing Hundreds Of Twitter Users Believed To Be 4chan Trolls Posing As Feminists*. BuzzFeed News. https://www.buzzfeednews.com/article/ryanhatesthis/your-slip-is-showing-4chan-trolls-operation-lollipop

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, *108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Collins, T. P., Crawford, J. T., & Brandt, M. J. (2017). No Evidence for Ideological Asymmetry in Dissonance Avoidance. *Social Psychology*, *48*(3), 123–134. https://doi.org/10.1027/1864-9335/a000300

Crawford, J. T. (2014). Ideological symmetries and asymmetries in political intolerance and prejudice toward political activist groups. *Journal of Experimental Social Psychology*, *55*, 284–298. https://doi.org/10.1016/j.jesp.2014.08.002

Daniels, G. R. (2009). Voter Deception. *Indiana Law Review*, *43*, 343–388.

De keersmaecker, J., & Roets, A. (2019). Is there an ideological asymmetry in the moral approval of spreading misinformation by politicians? *Personality and Individual Differences*, *143*, 165–169. https://doi.org/10.1016/j.paid.2019.02.003

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2018). *The Tactics & Tropes of the Internet Research Agency*. New Knowledge. https://cdn2.hubspot.net/hubfs/4326998/ira-report-rebrand_FinalJ14.pdf

Dobson, K., & Knezevic, I. (2018). “Ain’t Nobody Got Time for That!”: Framing and Stereotyping in Legacy and Social Media. *Canadian Journal of Communication; Toronto*, *43*(3), 381–397. http://dx.doi.org/10.22230/cjc.2018v43n3a3378

Faris, R. M., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). *Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election*. Berkman Klein Center for Internet & Society. https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf

Forelle, M., Howard, P., Monroy-Hernández, A., & Savage, S. (2015). Political Bots and the Manipulation of Public Opinion in Venezuela. *ArXiv:1507.07109 [Physics]*. http://arxiv.org/abs/1507.07109

Freelon, D. (2019). *Tweeting left, right, & center: How users and attention are distributed across Twitter* (pp. 1–38). John S. & James L. Knight Foundation. https://knightfoundation.org/reports/tweeting-left-right-center-how-users-and-attention-are-distributed-across-twitter/

Freelon, D., & Lokot, T. (2020). Russian Twitter disinformation campaigns reach across the American political spectrum. *Misinformation Review*, *1*(1).

Freelon, D., & Wells, C. (in press). Disinformation as Political Communication. *Political Communication*.

Green, J. L. (2006). *Digital Blackface: The Repackaging of the Black Masculine Image* [Miami University]. https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:miami1154371043

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Grossmann, M., & Hopkins, D. A. (2015). Ideological Republicans and Group Interest Democrats: The Asymmetry of American Party Politics. *Perspectives on Politics*, *13*(1), 119–139. https://doi.org/10.1017/S1537592714003168

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Hegelich, S., & Janetzko, D. (2016, March 31). Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. *Tenth International AAAI Conference on Web and Social Media*. Tenth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015

High Level Expert Group on Fake News and Disinformation. (2018). *A multi-dimensional approach to disinformation: Report of the independent High Level Group on fake news and online disinformation*. European Commission. https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

Hjorth, F., & Adler-Nissen, R. (2019). Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences. *Journal of Communication*, *69*(2), 168–192. https://doi.org/10.1093/joc/jqz006

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & Francois, C. (2018). *The IRA, Social Media and Political Polarization in the United States, 2012-2018*. University of Oxford. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf

Jamieson, K. H. (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press.

Jost, J. T., Stern, C., Rule, N. O., & Sterling, J. (2017). The Politics of Fear: Is There an Ideological Asymmetry in Existential Motivation? *Social Cognition*, *35*(4), 324–353. https://doi.org/10.1521/soco.2017.35.4.324

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83. https://doi.org/10.1016/j.copsyc.2018.01.003

Linvill, D., & Warren, P. L. (2018). *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*. http://pwarren.people.clemson.edu/Linvill_Warren_TrollFactory.pdf

Mueller, R. S. (2019). *The Mueller Report: Report on the Investigation into Russian Interference in the 2016 Presidential Election*. United States Department of Justice. https://www.justice.gov/storage/report.pdf

Peng, J., Detchon, S., Choo, K.-K. R., & Ashman, H. (2017). Astroturfing detection in social media: A binary n-gram–based approach. *Concurrency and Computation: Practice and Experience*, *29*(17).

Perrin, A., & Anderson, M. (2019, April 10). Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center*. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Popken, B. (2018, February 14). Twitter deleted Russian troll tweets. So we published more than 200,000 of them. *NBC News*. https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: Mapping the spread of astroturf in microblog streams. *Proceedings of the 20th International Conference Companion on World Wide Web*, 249–252. https://doi.org/10.1145/1963192.1963301

Robertson, A., Magdy, W., & Goldwater, S. (2018, June 15). Self-Representation on Twitter Using Emoji Skin Color Modifiers. *Twelfth International AAAI Conference on Web and Social Media*. Twelfth International AAAI Conference on Web and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17833

Roman, M. L. (2019). *Opposing Jim Crow: African Americans and the Soviet Indictment of U.S. Racism, 1928-1937*. U of Nebraska Press.

Singh, M., Bansal, D., & Sofat, S. (2016). Followers or Fradulents? An Analysis and Classification of Twitter Followers Market Merchants. *Cybernetics and Systems*, *47*(8), 674–689. https://doi.org/10.1080/01969722.2016.1237227

Stark, L. (2018). Facial recognition, emotion and race in animated social media. *First Monday*, *23*(9). https://doi.org/10.5210/fm.v23i9.9406

Stringhini, G., Wang, G., Egele, M., Kruegel, C., Vigna, G., Zheng, H., & Zhao, B. Y. (2013). Follow the Green: Growth and Dynamics in Twitter Follower Markets. *Proceedings of the 2013 Conference on Internet Measurement Conference*, 163–176. https://doi.org/10.1145/2504730.2504731

Twitter. (n.d.). *The Twitter Rules*. Retrieved May 15, 2019, from https://help.twitter.com/en/rules-and-policies/twitter-rules

Vargo, C. J., Guo, L., & Amazeen, M. A. (2017). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 1461444817712086. https://doi.org/10.1177/1461444817712086

Weedon, J., Nuland, W., & Stamos, A. (2017). *Information Operations and Facebook* (pp. 1–13). Facebook. https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

Weill, K. (2019, February 8). Pro-Trump Trolls Are Impersonating Black People on Twitter. *The Daily Beast*. https://www.thedailybeast.com/digital-blackface-pro-trump-trolls-are-impersonating-black-people-on-twitter-9

Wojcik, S., & Hughes, A. (2019). *How Twitter Users Compare to the General Public*. Pew Research Center. https://www.pewinternet.org/2019/04/24/sizing-up-twitter-users/

Table 1: Linvill and Warren's IRA account categories with example accounts (plus Black Troll)

Category	Assumed identity	Examples
Right Troll	White, far-right Americans	@foundingson @southlonestar @ten_gop
Left Troll	Far-left Americans of diverse backgrounds	@lgbtunitedcom @muslims_in_usa @newyorkdem
Black Troll	Black (left-wing) activists	@black4unity @gloed_up @trayneshacole
News Feed	American news outlets and aggregators	@dailylosangeles @onlinecleveland @richmondvoice
Hashtag Gamer	Players of "hashtag games," various identities, often apolitical	@andyhashtagger @fameonyoubitch @worldofhashtags
Fearmonger	Distributors of false information	@itsrealrobert @never2muchbass @originofkaty
Non-English	Accounts in other languages	@1488reasons @novostiputin @nyan_meow_meow
Unknown	Accounts with insufficient activity to classify them	@alwayshungrybae @ibetyouwill @tedcoolashell

Table 2: Negative binomial regression models predicting IRA retweets

	Model set 1: Black Trolls separated			Model set 2: Black Trolls combined		
Predictor	**Mean coef.**	**Coef. CI**	**Mean SE**	**Mean coef.**	**Coef. CI**	**Mean SE**
(Intercept)	-0.311	(-0.251, -0.37)	0.498	-0.931	(-0.87, -0.993)	0.498
BlackTroll	2.864***	(2.911, 2.818)	0.373	NA	NA	NA
LeftTroll	2.102**	(2.146, 2.057)	0.37	2.822***	(2.865, 2.778)	0.359
RightTroll	1.769*	(1.814, 1.724)	0.374	1.875*	(1.919, 1.83)	0.363
HashtagGamer	1.691*	(1.735, 1.647)	0.374	1.721*	(1.764, 1.678)	0.359
Commercial	0.794	(0.86, 0.729)	0.602	0.909	(0.972, 0.846)	0.588
NewsFeed	0.843	(0.889, 0.798)	0.373	0.888	(0.93, 0.845)	0.358
NonEnglish	2.551***	(2.598, 2.504)	0.38	2.548**	(2.592, 2.504)	0.363
Unknown	2.718***	(2.762, 2.674)	0.38	2.729***	(2.771, 2.686)	0.364
hashtag count	-0.055	(-0.049, -0.062)	0.059	-0.074	(-0.068, -0.081)	0.062
mention count	-0.318*	(-0.308, -0.328)	0.078	-0.296*	(-0.286, -0.305)	0.078
URL count	-0.314	(-0.303, -0.326)	0.109	-0.431*	(-0.418, -0.443)	0.114
tweet age	-0.002**	(-0.002, -0.002)	0	-0.002**	(-0.002, -0.002)	0
account age	-0.001*	(-0.001, -0.001)	0	-0.001*	(-0.001, -0.001)	0
follower count	0***	(0, 0)	0	0***	(0, 0)	0
following count	0**	(0, 0)	0	0**	(0, 0)	0
tweet length	0.01**	(0.01, 0.01)	0.001	0.01**	(0.01, 0.01)	0.002
contains image	1.681***	(1.696, 1.667)	0.124	1.725***	(1.74, 1.71)	0.13
contains video	2.118***	(2.16, 2.075)	0.333	2.217**	(2.263, 2.172)	0.358
Mean BIC	9663.374			8915.816		

All coefficients are unstandardized.
* = $p < .05$; ** = $p < .01$; *** = $p < .001$.
The starred *p* thresholds are based on the upper limit of the 95% confidence interval for each coefficient's mean *p* value.

Table 3: Negative binomial regression models predicting IRA likes

	Model set 1: Black trolls separated			Model set 2: Black trolls combined		
Predictor	**Mean coef.**	**Coef. CI**	**Mean SE**	**Mean coef.**	**Coef. CI**	**Mean SE**
(Intercept)	1.614	(1.664, 1.565)	0.506	1.319	(1.372, 1.265)	0.497
BlackTroll	2.812***	(2.845, 2.78)	0.411	NA	NA	NA
LeftTroll	2.26***	(2.291, 2.229)	0.41	2.784***	(2.814, 2.754)	0.386
RightTroll	1.653**	(1.686, 1.621)	0.413	1.733**	(1.764, 1.701)	0.391
HashtagGamer	2.304***	(2.335, 2.272)	0.409	2.245***	(2.275, 2.216)	0.383
Commercial	2.155*	(2.2, 2.11)	0.557	2.285*	(2.328, 2.242)	0.538
NewsFeed	0.554	(0.585, 0.522)	0.415	0.615	(0.646, 0.584)	0.39
NonEnglish	1.26	(1.297, 1.223)	0.435	1.296*	(1.331, 1.261)	0.408
Unknown	1.125	(1.161, 1.089)	0.465	1.123	(1.157, 1.088)	0.436
hashtag count	-0.064	(-0.058, -0.069)	0.056	-0.083	(-0.077, -0.089)	0.06
mention count	-0.191	(-0.185, -0.197)	0.064	-0.17	(-0.164, -0.176)	0.063
URL count	-1.055***	(-1.042, -1.067)	0.108	-1.262***	(-1.248, -1.276)	0.114
tweet age	-0.003***	(-0.003, -0.003)	0	-0.003***	(-0.003, -0.003)	0
account age	-0.001**	(-0.001, -0.001)	0	-0.001*	(-0.001, -0.001)	0
follower count	0***	(0, 0)	0	0***	(0, 0)	0
following count	0*	(0, 0)	0	0*	(0, 0)	0
tweet length	0.007*	(0.007, 0.007)	0.001	0.007*	(0.007, 0.007)	0.001
contains image	1.52***	(1.534, 1.506)	0.108	1.554***	(1.568, 1.539)	0.113
contains video	1.811**	(1.854, 1.767)	0.283	1.847**	(1.895, 1.8)	0.299
Mean BIC	9364.267			8535.074		

All coefficients are unstandardized.
* = $p < .05$; ** = $p < .01$; *** = $p < .001$.
The starred *p* thresholds are based on the upper limit of the 95% confidence interval for each coefficient's mean *p* value.

Table 4: Negative binomial regression models predicting IRA replies

	Model set 1: Black trolls separated			Model set 2: Black trolls combined		
Predictor	**Mean coef.**	**Coef. CI**	**Mean SE**	**Mean coef.**	**Coef. CI**	**Mean SE**
(Intercept)	-2.297	(-2.245, -2.349)	0.863	-2.735*	(-2.678, -2.792)	0.859
BlackTroll	2.51**	(2.547, 2.473)	0.793	NA	NA	NA
LeftTroll	2.314**	(2.351, 2.276)	0.798	2.684**	(2.722, 2.646)	0.78
RightTroll	2.254**	(2.292, 2.216)	0.795	2.487**	(2.525, 2.449)	0.782
HashtagGamer	1.652	(1.69, 1.614)	0.806	1.782	(1.821, 1.742)	0.79
Commercial	1.826	(1.878, 1.774)	0.945	2.073	(2.125, 2.021)	0.927
NewsFeed	0.411	(0.451, 0.37)	0.82	0.591	(0.634, 0.549)	0.803
NonEnglish	2.285*	(2.324, 2.246)	0.806	2.427**	(2.468, 2.387)	0.789
Unknown	1.83	(1.872, 1.788)	0.842	1.928	(1.971, 1.885)	0.821
hashtag count	-0.23*	(-0.224, -0.235)	0.07	-0.254*	(-0.248, -0.26)	0.073
mention count	0.016	(0.022, 0.009)	0.081	0.022	(0.028, 0.017)	0.079
URL count	-0.915**	(-0.902, -0.929)	0.144	-1.049***	(-1.034, -1.064)	0.15
tweet age	-0.002***	(-0.002, -0.002)	0	-0.002***	(-0.002, -0.002)	0
account age	-0.001	(-0.001, -0.001)	0	0	(0, 0)	0
follower count	0***	(0, 0)	0	0***	(0, 0)	0
following count	0*	(0, 0)	0	0*	(0, 0)	0
tweet length	0.009**	(0.009, 0.009)	0.002	0.009*	(0.01, 0.009)	0.002
contains image	0.893**	(0.907, 0.879)	0.136	0.983**	(0.996, 0.969)	0.14
contains video	1.456*	(1.49, 1.422)	0.3	1.478*	(1.511, 1.445)	0.314
Mean BIC	3962.978			3640.595		

All coefficients are unstandardized.
* = $p < .05$; ** = $p < .01$; *** = $p < .001$.
The starred *p* thresholds are based on the upper limit of the 95% confidence interval for each coefficient's mean *p* value.

Table 5: False amplification rates for retweets and replies by account type

Account type	Retweet false amplification rate (%)	Reply false amplification rate (%)
BlackTroll	0.155	0.203
LeftTroll	0.076	0.065
RightTroll	0.090	0.139
HashtagGamer	29.628	1.359
Commercial	0.000	0.000
NewsFeed	5.366	11.106
NonEnglish	7.490	2.913
Unknown	63.205	4.007
Fearmonger	73.321	6.202

Figure 1: Number of unique users per IRA category

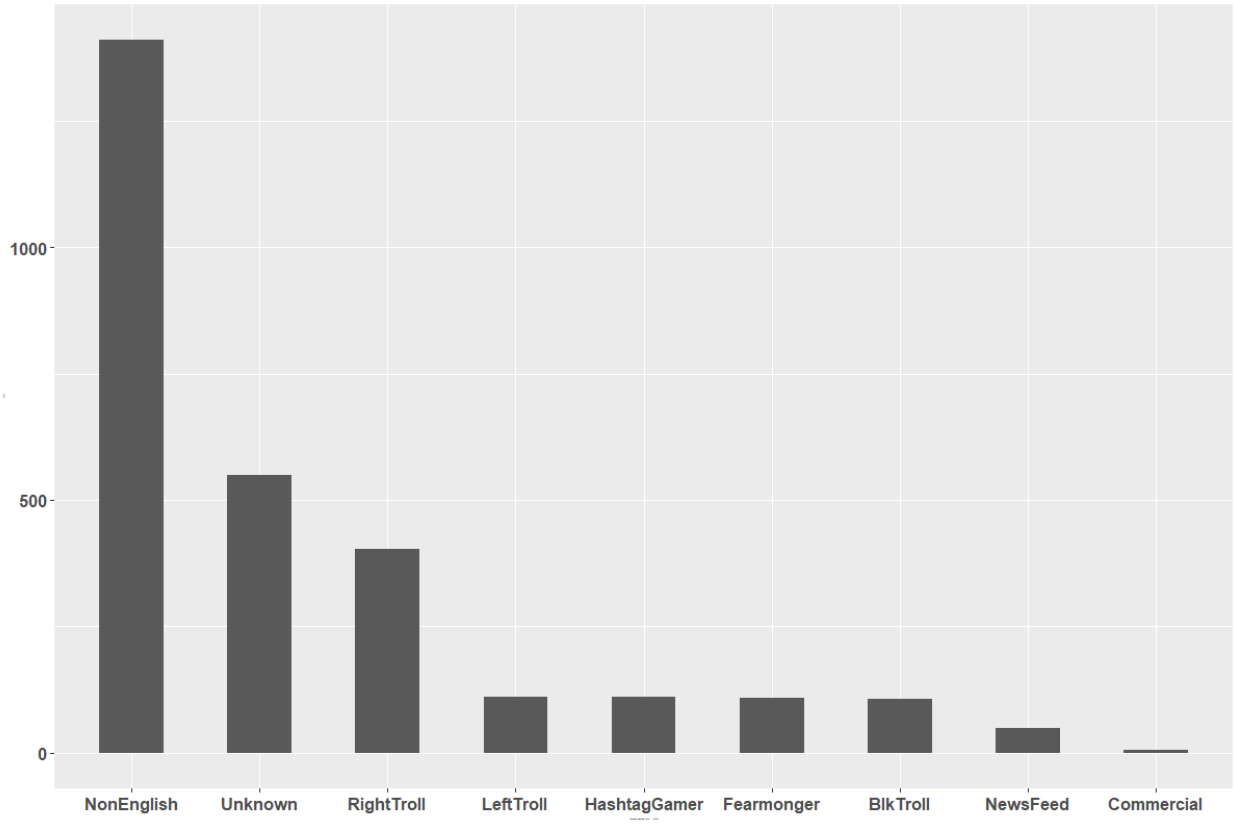


Figure 2: Number of tweets per IRA category

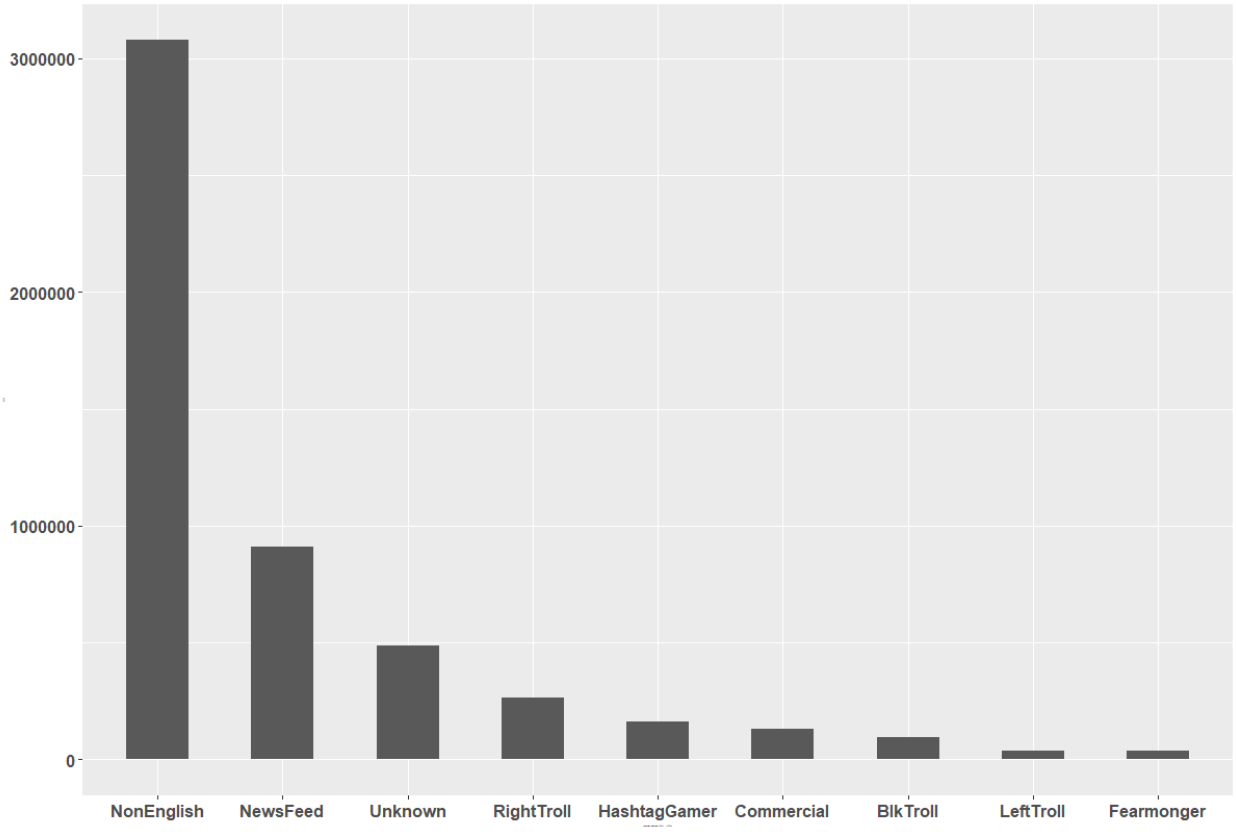
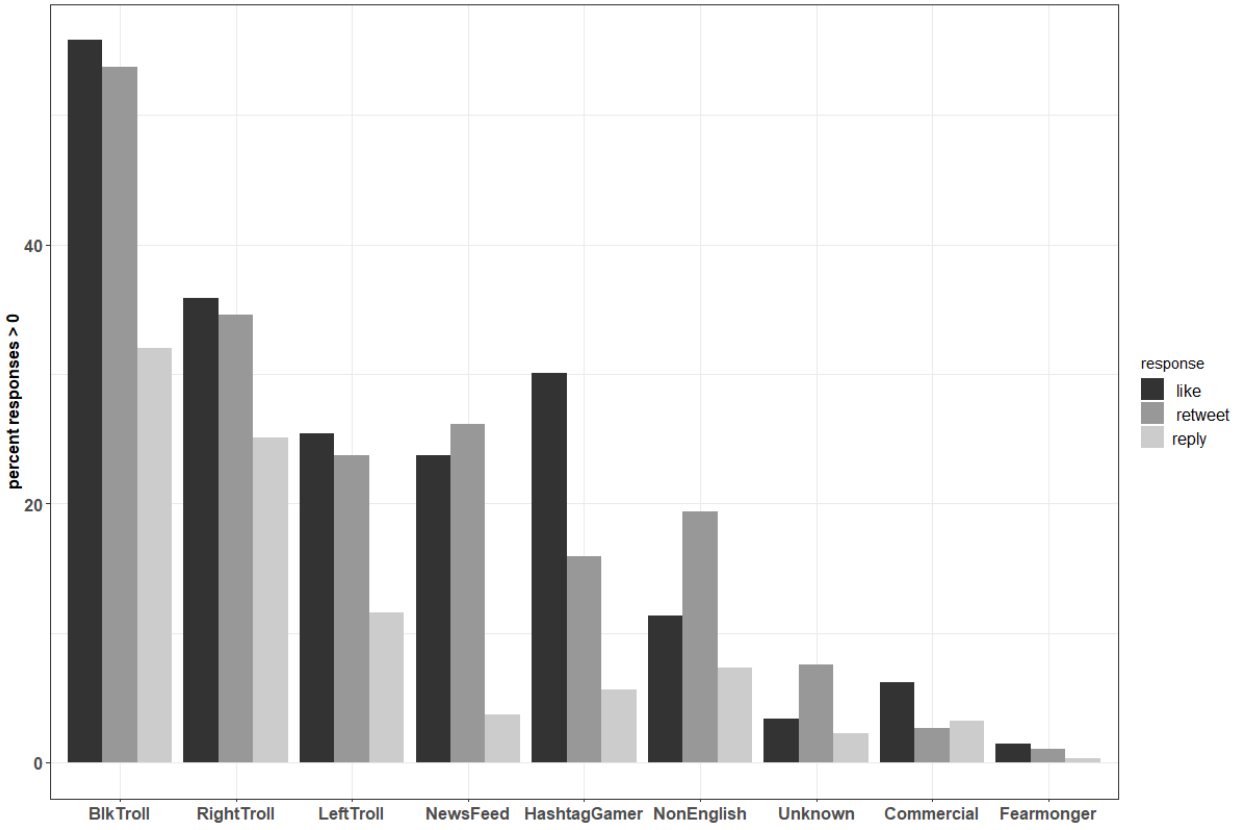


Figure 3: Percentages of IRA tweets attracting at least one of each response type by account type



Notes

[2] The redacted dataset contains tweets from 188 accounts that are absent from in the unredacted dataset for unknown reasons. As a result, the latter dataset contains roughly 3% fewer tweets than the former.

[3] We decided to use Linvill and Warren's category scheme rather than our own because theirs is based on their IRA dataset of over three million tweets; while ours was based on the NBC dataset, which contains slightly over 200,000 tweets.

[4] We did not search for Black right-leaning trolls because Black conservatives and Republicans are extremely rare in US politics (Bonn, 2019), and none of the preliminary work or theory on the IRA indicated they created any Black conservative sockpuppets.

[5] We chose Fearmonger as the reference because it attracted the fewest of all three responses, and thus can serve as a baseline from which to measure the relative strength of the other categories' coefficients.

[6] This analysis was not possible for likes due to a lack of data concerning individual accounts' liking behaviors.