Bots Amplify and Redirect Hate Speech in Online Discourse About Racism During the COVID-19 Pandemic

Joshua Uyheng, Daniele Bellutta, and Kathleen M. Carley

Abstract
Online talk about racism has been salient throughout the COVID-19 pandemic. Yet while such social media conversations reflect existing tensions in the offline world, the same discourse has also become a target for information operations aiming to heighten social divisions. This article examines Twitter discussions of racism in the first and sixth months since COVID-19 was accorded pandemic status by the World Health Organization and uncovers dynamic associations with bot activity and hate speech. Humans initially constituted the most hateful accounts in online conversations about racism in March, but in August, bots dominated hate speech. Over time, greater bot activity likewise amplified levels of hate speech a week later. Moreover, while discourse about racism in March primarily featured an organic focus on racial identities like Asian and Chinese, we further observed a bot-dominated focus in August toward political identities like president, Democrat, and Republican. Although hate speech targeting Asian groups remained present among racism discussions in August, these findings suggest a bot-fueled redirection from focusing on racial groups at the onset of the pandemic to targeting politics closer to the 2020 US elections. This work enhances understanding of the complexity of racism discussions during the pandemic, its vulnerability to manipulation through information operations, and the large-scale quantitative study of inorganic hate campaigns in online social networks.

Keywords
racism, hate speech, bots, social media, COVID-19 pandemic

Introduction
The COVID-19 pandemic has not only triggered historic public health crises but also exacerbated various social conflicts (Chiriboga et al., 2020; Van Bavel et al., 2020). Race and racism have featured prominently as a major axis of contention throughout the pandemic, especially as Asian populations in particular became hatefully associated with the pathogen and blamed for the global outbreak (Egede & Walker, 2020; Gee et al., 2020; Reny & Barreto, 2022). In a time of greater concern about nationalism and xenophobia around the world, such racial divisions have not only flared up organically among people but have also been mobilized by political leaders (Chan & Montt Strabucchi, 2021; Daniels et al., 2021; Dhanani & Franz, 2021; Kim & Kesari, 2021). These social forces have contributed to severe, disproportionate impacts for racial and ethnic minorities during the pandemic, ranging from harmful interpersonal encounters to systemic disparities in pandemic outcomes (Devakumar et al., 2020; Gover et al., 2020; Khazanchi et al., 2020; Le et al., 2020; Li & Galea, 2020; Tan et al., 2022).

These issues play out in important ways in cyberspace, where “infodemics” have spread alongside COVID-19 outbreaks in the offline world (Gallotti et al., 2020). In this context, racism during the pandemic has been a major topic of discussion in various online spaces. Although recent scholarship indicates that some of these discussions have facilitated community resilience among Asian populations targeted by pandemic racism (Abidin & Zeng, 2020; Buerger, 2021; Kuo et al., 2020), social media has also served as a significant platform for the proliferation of racist...
rhetoric and xenophobic discourse (Abd-Alrazaq et al., 2020; Dubey, 2020; Stechemesser et al., 2020). These findings emphasize growing consensus around the assessment that social media constitutes a major domain in which public health crises play out, especially as regards their distinctly social dimensions (Islam et al., 2020; Tsao et al., 2021; Van Bavel et al., 2020).

Against this backdrop of crisis-fueled social conflict, online discourse about racism—including both racist rhetoric and counter-discourse responding to it—constitutes a ripe target for information operations, such as those driven by automated social bots (Arif et al., 2018; K. M. Carley, 2020; Ferrara et al., 2016). Yet while a wealth of research documents the prevalence of online racism discussion during the pandemic, evidence around its potential manipulation is relatively scant. There is ample scientific knowledge of bot activity in online talk about the COVID-19 pandemic, especially regarding their spread of COVID-19 conspiracy theories (Ferrara, 2020; Moffitt et al., 2021) and more general low-credibility information (Himelein-Wachowiak et al., 2021; Xu & Sasahara, 2022). However, as Kim and Kesari (2021) point out, much of the literature on hate and misinformation is relatively disconnected, even when the two phenomena are closely linked.

This study addresses these gaps by explicitly considering the joint roles of hate speech and bot activity in online discussions of racism during the COVID-19 pandemic. We propose a computational social science framework applicable at scale to characterize how bots may have influenced online discussions of racism by injecting hatred into public discourse and by reframing how these hateful sentiments were targeted (Heise, 1987; Joseph et al., 2016; Uyeheng & Carley, 2020, 2021a). We investigate tweets in the racism discussion that are infused with hate speech and potentially contain messages which: (a) express explicitly racist rhetoric targeting minoritized racial groups or (b) hatefully use “racist” as an epithet against individuals or groups. In probing these dynamics, our aim is not to obscure the role of organic conflicts around the topic of racism, but rather to show how precisely these extant divisions may be mobilized and manipulated through bot-driven information operations (Arif et al., 2018; K. M. Carley, 2020; Starbird, 2019).

To investigate these processes, we focused on two points in time marking the first 6 months since the World Health Organization declared COVID-19 to be a pandemic: March 2020 and August 2020. At these time points, we show that even as racism comprised (and still comprises) a core topic of human conversations during the pandemic, particularly concerned with anti-Asian sentiments, pandemic discourse soon became entangled in the politicized activity of social bots. Organic rhetoric hatefully targeting Asian groups of people remained endemic throughout our data, yet over time, our results also demonstrate an inorganic, bot-fueled amplification of hate speech featuring a discursive pivot toward political conflict. As we discuss in our concluding sections, these findings underscore the racialized dimensions of misinformation and disinformation in and beyond the pandemic (Frellon & Wells, 2020; Kuo & Marwick, 2021; Reddi et al., 2021) and the necessity of carrying these insights forward to tackle the ongoing global crisis (Chiriboga et al., 2020; Kim & Kesari, 2021).

In sum, this article therefore poses the following research questions:

- **RQ1**: How did bots affect the levels of hate speech in online racism discourse during the pandemic?
- **RQ2**: How did bots affect the targets of hate speech in online racism discourse during the pandemic?

### Related Work

**Online Discussions of Pandemic Racism**

Social science scholarship posits a tight coupling between crises and the deepening of social divides, suggesting that the salience of race and racism during the COVID-19 pandemic is not a surprising phenomenon (Clissold et al., 2020; Elias et al., 2021). Chan and Montt Strabucchi (2021), for instance, show how the pandemic activates deeply held stereotypes about Asian populations across a range of political and cultural discourses by government leaders. Daniels et al. (2021) and Dhanani and Franz (2021) independently verify through experiments that reminders of the pandemic stoke more negative attitudes toward racial out-groups and prompt prioritization of resources for racial in-groups. Moreover, these effects are not constrained to the cognitive and affective domain; they have precipitated large-scale harmful behaviors toward racial minorities. Le et al. (2020) document the negative ways associations with the pathogen impact everyday interpersonal encounters for Asian populations. Gover et al. (2020), meanwhile, pointedly highlight an uptick in Asian-directed hate crimes throughout the duration of the pandemic. More broadly, a wealth of public health research indicates that racial differences in socioeconomic status are interwoven into the ways the pandemic is experienced—encompassing not just Asian populations but all minority racial and ethnic groups—as racial disparities abound in case and fatality rates, resource allocation, and policy outcomes around the world (Devakumar et al., 2020; Gee et al., 2020; Khazanchi et al., 2020; Tan et al., 2022).

From this standpoint, the role of social media at the nexus of racism and the pandemic is a multifaceted one. Much has been said regarding the need to investigate the public health implications of the flood of online information that has overwhelmed societies in a time of crisis (Gallotti et al., 2020; Islam et al., 2020). As burgeoning scholarship indicates, digital platforms may facilitate positive outcomes. As Abidin and Zeng (2020) and Kuo et al. (2020) note, for instance, online discussions of racism can facilitate community-building and shared coping against discrimination in the offline
world. Buerger (2021) illustrates, too, how online collectives may actively counter racist rhetoric in digital spaces through moderation practices to sustain improved online discourse.

However, at the same time, studies also urgently sound the alarm around the negative impacts of social media vis-à-vis racism discourse during the pandemic. In an infoveillance study of the top concerns of Twitter users during the pandemic, Abd-Alrazaq et al. (2020) discover that racist associations between the virus and China are a dominant topic alongside public health issues. Dubey (2020) similarly shows that the most negative emotions expressed in online talk about the pandemic surround references to Asian populations. Stechemesser et al. (2020) observe that at the start of the pandemic, not only was there a surge in explicitly expressed racist rhetoric against Asians online, these messages also received a wave of “likes” signaling support for these sentiments. These effects furthermore increase with the exponential growth in infections worldwide, highlighting long-standing theoretical links between crises and conflict (Stechemesser et al. 2020).

**Bots and the COVID-19 Infodemic**

Especially in recent years, such widespread contentions in the digital sphere do not often erupt entirely organically but are instead subject to a variety of inorganic influences (Freelon & Wells, 2020; Starbird, 2019). Automated accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2016). Due to their automated nature, social bots may be coordinated to engage in “computational propaganda” through large-scale information operations to influence public discourse (Beskow & Carley, 2019; Woolley & Howard, 2016). K. M. Carley (2020) succinctly summarizes such actions as narrative maneuvers, or actions which positively or negatively impact the framing of an issue; and network maneuvers, or actions which positively or negatively impact information flow in large-scale conversations, such as who talks to whom and receives information from each other.

A range of such bot activities have been documented in the context of the COVID-19 pandemic. Ferrara (2020), for instance, highlights the outsize role played by bots in the spread of conspiracy theories early in the pandemic, especially those linked to QAnon and elaborate narratives of the coronavirus as originating from a Wuhan lab. Moffitt et al. (2021) reaffirm these findings, showing that Twitter accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2019). Automated accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2016). Due to their automated nature, social bots may be coordinated to engage in “computational propaganda” through large-scale information operations to influence public discourse (Beskow & Carley, 2019; Woolley & Howard, 2016). K. M. Carley (2020) succinctly summarizes such actions as narrative maneuvers, or actions which positively or negatively impact the framing of an issue; and network maneuvers, or actions which positively or negatively impact information flow in large-scale conversations, such as who talks to whom and receives information from each other.

A range of such bot activities have been documented in the context of the COVID-19 pandemic. Ferrara (2020), for instance, highlights the outsize role played by bots in the spread of conspiracy theories early in the pandemic, especially those linked to QAnon and elaborate narratives of the coronavirus as originating from a Wuhan lab. Moffitt et al. (2021) reaffirm these findings, showing that Twitter accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2019). Automated accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2016). Due to their automated nature, social bots may be coordinated to engage in “computational propaganda” through large-scale information operations to influence public discourse (Beskow & Carley, 2019; Woolley & Howard, 2016). K. M. Carley (2020) succinctly summarizes such actions as narrative maneuvers, or actions which positively or negatively impact the framing of an issue; and network maneuvers, or actions which positively or negatively impact information flow in large-scale conversations, such as who talks to whom and receives information from each other.

A range of such bot activities have been documented in the context of the COVID-19 pandemic. Ferrara (2020), for instance, highlights the outsize role played by bots in the spread of conspiracy theories early in the pandemic, especially those linked to QAnon and elaborate narratives of the coronavirus as originating from a Wuhan lab. Moffitt et al. (2021) reaffirm these findings, showing that Twitter accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2019). Automated accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2016). Due to their automated nature, social bots may be coordinated to engage in “computational propaganda” through large-scale information operations to influence public discourse (Beskow & Carley, 2019; Woolley & Howard, 2016). K. M. Carley (2020) succinctly summarizes such actions as narrative maneuvers, or actions which positively or negatively impact the framing of an issue; and network maneuvers, or actions which positively or negatively impact information flow in large-scale conversations, such as who talks to whom and receives information from each other.

A range of such bot activities have been documented in the context of the COVID-19 pandemic. Ferrara (2020), for instance, highlights the outsize role played by bots in the spread of conspiracy theories early in the pandemic, especially those linked to QAnon and elaborate narratives of the coronavirus as originating from a Wuhan lab. Moffitt et al. (2021) reaffirm these findings, showing that Twitter accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2019). Automated accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2016). Due to their automated nature, social bots may be coordinated to engage in “computational propaganda” through large-scale information operations to influence public discourse (Beskow & Carley, 2019; Woolley & Howard, 2016). K. M. Carley (2020) succinctly summarizes such actions as narrative maneuvers, or actions which positively or negatively impact the framing of an issue; and network maneuvers, or actions which positively or negatively impact information flow in large-scale conversations, such as who talks to whom and receives information from each other.

A range of such bot activities have been documented in the context of the COVID-19 pandemic. Ferrara (2020), for instance, highlights the outsize role played by bots in the spread of conspiracy theories early in the pandemic, especially those linked to QAnon and elaborate narratives of the coronavirus as originating from a Wuhan lab. Moffitt et al. (2021) reaffirm these findings, showing that Twitter accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2019). Automated accounts, also known as social bots, have played a significant role in this context (Ferrara et al., 2016). Due to their automated nature, social bots may be coordinated to engage in “computational propaganda” through large-scale information operations to influence public discourse (Beskow & Carley, 2019; Woolley & Howard, 2016). K. M. Carley (2020) succinctly summarizes such actions as narrative maneuvers, or actions which positively or negatively impact the framing of an issue; and network maneuvers, or actions which positively or negatively impact information flow in large-scale conversations, such as who talks to whom and receives information from each other.

Framing Hate Through an Identity Lens

Canonically defined as abusive language targeting a social group (Bluel et al., 2018; Davidson et al., 2017; MacAvaney et al., 2019), hate speech has long represented a central problem for social media even prior to the pandemic. In a cross- national survey, Reichelmann et al. (2021) observe that nearly 90% of respondents have observed or been exposed to online hate on social media. Especially during the pandemic, however, mounting evidence of increased hate speech—particularly in the form of racist rhetoric—bears unique significance given associations between online hate and violent behavior in the offline world (Van Bavel et al., 2020; Weber et al., 2020). Hoover et al. (2021), for instance, show that prejudicial expressions online are most likely in locations with specific configurations of moral values. Awan and Zempi (2016) similarly show that anti-Muslim hate crimes and online experiences of hate are interlinked. Evidence suggests that these online–offline links are bidirectional, as online hate reflects offline contexts of discrimination Lozada et al. (2021), while online hate simialrly potentially births new perpetrators offline in a “cycle of violence” (Wachs et al., 2022).

Our approach to hate speech—especially as manipulated by social bots in the context of the online racism discussion—is complicated by its inherent complexity as a social construct. Pohjonen and Udupa (2017), for instance, argue that what counts as “hate” may vary across cultures, where use of social media may be governed by unique conceptions of “extreme” speech defying binary definitions of hate versus non-hate. In this context, one way we precisely organize our conception of online hate falls in line with new theorizing around “identity propaganda” as advanced by Reddi...
et al. (2021). Through this lens, we consider hate speech as a specific form of abusive discourse which also “exploits social orders as communication strategically designed to undermine or manipulate target populations in pursuance of a political goal through appeals about identity or identities that accord with racial and other power structures” (Reddi et al. 2021, p. 5).

Viewing hate through an identity lens finds purchase not only in recent theoretical developments but also in methodological and empirical ones. A focus on identities opens up possibilities for large-scale analysis of online hate by harnessing prior work on measuring identities in text (Heise, 1987; Joseph et al., 2016). Recent work shows that across a wide range of benchmark datasets (Chetty & Alathur, 2018; Davidson et al., 2017; Fortuna & Nunes, 2018), abusive language consistently co-occurs with terms referencing identity categories in known instances of hate speech, and these lexical patterns may be harried for accurate and interpretable hate speech detection across languages and social media platforms (Uyheng & Carley, 2021b). Although such computational methods are not without their own limitations, they nonetheless introduce a systematic framework by which hate speech in online racism discussions can be identified (i.e., more or less likely to be hateful), as well as characterized (i.e., what identities are targeted). Orienting our approach to hate in this manner does not capture the complexity of online hate in its entirety (Pohjonen & Udupa, 2017); however, it may still meaningfully surface the social dimensions of hate beyond traditionally dichotomous classification schemes (Reddi et al., 2021) and thus provides a scalable, practical framework by which to specify how bots engage in hateful behavior in the context of online racism discourse and potentially introduce discursive shifts into the broader conversation (K. M. Carley, 2020; Uyheng et al., 2020).

**Methods**

**Data**

To examine the relationships between hate speech and bot activity in online discussion of racism during the pandemic, we analyzed a large-scale Twitter dataset collected by Huang and Carley (2020). This original dataset consisted of over 200 million pandemic-related tweets obtained using the Twitter streaming API with search terms related to COVID-19.

To tailor this work for our analysis of online racism discussions, we first took a subset of this large dataset consisting of tweets that mentioned racism-related terms such as “racist,” “bigot,” and “xenophobic.” We refer to this first dataset as **Tweets Mentioning Racism**. To detect temporal changes in this discourse, we obtained a set of Tweets Mentioning Racism from March and another set from August. Given insights from past work (Abidin & Zeng, 2020; Buerger, 2021), we did not assume a priori that these tweets were all necessarily hateful; rather, we acknowledged the meaningful possibility that these tweets may in fact be calling out racism—either in general or in response to another Twitter account.

For this reason, we collected a second dataset which we refer to as **Discussion-Sparking Tweets**. This second dataset contained any tweets that had been replied to by Tweets Mentioning Racism and was obtained for both March and August. These datasets were therefore used to examine whether there was a “call-out” dynamic in which Tweets Mentioning Racism were accusing others of being racist. We consequently viewed online racism discourse as both a dynamic and an interactive phenomenon.

Table 1 summarizes basic statistics of each dataset in this study. Over both time periods, we used a multilingual list of racism-related terms to filter Tweets Mentioning Racism, which are summarized in Table 2. Using these filtered tweets, we also isolated those that were replying to other tweets and attempted to collect those original tweets. These tweets that had been replied to with mentions of racism then formed the sets of Discussion-Sparking Tweets.

It should be noted that, as with any study involving Twitter samples, the conclusions drawn from a particular dataset may not necessarily be representative of the entire conversation surrounding a subject on Twitter (Morstatter et al., 2013). In addition, hate speech may be under-represented in our sets of Discussion-Sparking Tweets. Although the Tweets Mentioning Racism were collected via a live Twitter data stream, the tweets being replied to with these mentions of racism were necessarily collected after the fact. This means that some tweets could have been removed before they could be collected. Our data collection obtained 84.60% of the original Discussion-Sparking Tweets from March 2020 as well as 81.88% of those in August. The missing tweets could have been deleted or made private. Furthermore, because of our data collection strategy, all of our conclusions only apply to Twitter conversations about racism, not racist speech in particular.

**Tools**

In this work, we relied on a series of existing computational social science tools to measure and predict levels of hate speech, bot activity, and the use of various identities in online discourse. In general, each of the following tools may be modified or extended through constantly evolving modeling techniques. However, for the purposes of this work, our analytical focus lay not in tool development per se but rather in designing and implementing a theory-based and problem-driven framework for analyzing their relationships with respect to the case study at hand. Our selection of existing tools thus constitutes a strength of the framework we propose, since its individual components may be flexibly adapted in line with ongoing methodological advancements (Uyheng et al., 2020). From a practical standpoint, deploying previously validated tools may also constitute a typical use
Hate Speech Detection. We detected hate speech in our datasets with a machine learning model trained on a seminal benchmark dataset of hate speech (Davidson et al., 2017). Using the commercial NetMapper software (Altman et al., 2017; K. M. Carley, 2014), we obtained for each tweet a feature set of multilingual lexical counts derived from the psycholinguistic literature (Pennebaker et al., 2003). The hate speech detection model, based on a random forest classifier, utilizes these lexical counts to generate a score ranging from 0 to 1, estimating the probability that the given tweet is hate speech. Prior research has shown that the model achieves a weighted F1 score of 83% on the benchmark data (Uyheng & Carley, 2020, 2021b). While more complex hate speech detection methods may certainly be explored (Fortuna & Nunes, 2018), our priorities were scalability and interpretability.

Identity Lexicon. Resonating with recent scholarship on identity propaganda as a framework for online harms (Reddi et al., 2021), we sought to characterize hate speech in online discourse about racism through the use of identity terms, the use of which we counted with the NetMapper software (Altman et al., 2017; K. M. Carley, 2014; L. R. Carley et al., 2018). The identity lexicon we employed relies on previously validated dictionaries derived from extensive cross-cultural survey research (Heise, 1987). Prior work had repurposed these existing dictionaries of identity terms for computational modeling of texts (Joseph et al., 2016). These same identity terms have also been further classified in past work as being related to gender (e.g., “woman,” “transgender”), politics (e.g., “president,” “senator”), race or ethnicity (e.g., “Asian,” “Black”), religion (e.g., “Muslim,” “Christian”), or other identity categories. We were particularly interested in tweets with high hate speech scores that also mentioned either these particular identity terms or their broader categories. This method had been previously used to empirically track the targets of online hate (Uyheng & Carley, 2020, 2021a).

Bot Detection. Finally, to detect bot activity, we used the BotHunter machine learning model (Beskow & Carley, 2018). BotHunter has been trained on a variety of known bot datasets and shown to be competitive with state-of-the-art models (Beskow, 2020). Prior research has also successfully applied the BotHunter model across diverse settings to characterize bot-driven information operations (King et al., 2020; Moffitt et al., 2021; Uyheng et al., 2020). While bot activity is a diverse and evolving phenomenon (Ferrara et al., 2016), the training of BotHunter on diverse datasets of known information operations makes it particularly effective at detecting automated behavior aimed at amplifying particular narratives or disrupting online conversations (Beskow, 2020). BotHunter produces a probabilistic score that quantifies the likelihood that a given Twitter account is a bot. In most analyses, we used the entire distribution of BotHunter scores to capture more complete hate speech dynamics as accounts become more or less likely to be bots. However, in instances where we employed a binary classification of bots versus humans, we used a threshold of a 70% probability that a given account is a bot. Because we used a conservatively high threshold, we may have underestimated the total number of bots in the conversation. However, this also yields higher precision across predictions and homes in on the most bot-like accounts.

Measuring Bot Impacts on Hate

Utilizing the outputs of the tools above, we measured the ways bots may have influenced the prevalence of hate speech in online discussions of racism. Although we do not
make strict causal claims due to the limitations of social media data (Morstatter et al., 2013), evidence for bot impacts can nonetheless be obtained through an assessment of account behavior and conversation dynamics over time (K. M. Carley, 2020).

**Account Level.** How much hate could be attributed to bots in the online conversation about racism? We assessed this using a multiple regression model that tested the variation of bot probabilities over hate speech scores while controlling for tweet type (Tweet Mentioning Racism or Discussion-Sparking Tweet) and month (March or August). Taken together, these determined whether a bulk of hateful messages reliably originated from bots. From an information operation standpoint, this shed light on the potential objectives of bots in online discourse about racism. The models were standardized so effects could be interpreted on the scale of variable standard deviations.

**Conversation Level.** Was bot activity associated with higher levels of hate speech in the broader conversation? This transitioned our analysis from individual bot activities to reflect their collective impacts on online discussion of racism as a whole. To quantify this, we used a path modeling framework to capture longitudinal effects (Byrne, 2005). A path model captures associations between variables while precisely quantifying indirect relationships such as mediation effects. For longitudinal analysis, our goal is to isolate the effect of current bot activity on future levels of hate speech.

To this end, we assessed daily average bot probabilities and average hate speech scores separated by a 1-week time lag, which was chosen to avoid the susceptibility to noise that could come with a lag of only a day. Using path analysis, we then estimated the effects of: (a) current bot scores on current hate speech scores, (b) future bot scores on future hate speech scores, (c) current bot scores on future bot scores, (d) current hate speech scores on future hate speech scores, (e) current bot scores on future hate speech scores, and (f) current hate speech scores on future bot scores. Effects (a) and (b) capture the concurrent activities of bots in injecting hate into the online conversation, effects (c) and (d) control for autocorrelation between current and future levels of bot activity and hate speech, while effects (e) and (f), respectively, capture crucial measures of whether bots introduce enduring shifts in levels of hate or merely respond to existing levels of hate.

The adequacy of the path model was assessed using standard fitness measures including the comparative fit index (CFI; lower-bound cutoff: 0.90), the Tucker–Lewis index (TLI; lower-bound cutoff: 0.90), the root mean square error of approximation (RMSEA; upper-bound cutoff: 0.10), and the standardized root mean square residual (SRMR; upper-bound cutoff: 0.10) (Byrne, 2005). All variables in the model included controls for month and tweet type as binary variables. The models were also standardized so effects could be interpreted on the scale of variable standard deviations.

**Identifying Bot Targets of Hate**

We were interested not only in the bot-driven prevalence of hate speech but also in their targets from an identity lens (Reddi et al., 2021). Harnessing an identity lexicon (Heise, 1987; Joseph et al., 2016), we measured identity targeting on coarse-grained and fine-grained levels.

**Coarse-Grained Targets.** Coarse-grained analysis measured the extent to which hateful tweets in online discussions of racism targeted identities across race/nationality, politics, gender, and religion. We estimated logistic regression models on the probability of mentioning at least one identity term belonging to each category, given the tweet’s hate speech score, whether it was tweeted by a bot at a 70% probability threshold, whether it was tweeted in March or August, and whether it was a Tweet Mentioning Racism or a Discussion-Sparking Tweet. Higher associations with bot-produced hate indicated which identities were hatefully targeted by bots; conversely, higher associations with human-produced hate pointed to identities hatefully targeted by human accounts.

**Fine-Grained Targets.** Fine-grained analysis examined particular identity terms mentioned in tweets by bots and humans. Here, we sought to answer two questions. First, did bots target some identities more than expected? Second, did bots increase or decrease their focus on some identities over time? We answered these questions separately for Tweets Mentioning Racism and Discussion-Sparking Tweets.

To answer the first question, we took the prevalence of bots in tweets containing every single identity term. We then compared these observed prevalence rates to the expected prevalence if bots were distributed uniformly across tweets. For example, if 20% of Discussion-Sparking Tweets in March were from bots, then the “expected” proportion of tweets mentioning “chinese” by bots would also be 20%. For each identity term, we performed a one-sided, one-sample proportion test to evaluate whether the actual proportion of tweets coming from bots was significantly higher than the expected proportion. To answer the second question, on the other hand, we performed two-sample proportion tests between time points. For instance, if “Chinese” was mentioned by bots in Tweets Mentioning Racism both in March and August, a two-sample proportion test would be run to test whether bots increased or decreased their rate of mentioning “Chinese” over time.

**Results**

Analysis of the online discussion of racism in the context of the COVID-19 pandemic revealed two key insights, which we unpack in the succeeding sections. First, bots amplified hate speech in online discourse about racism. Although humans were initially more hateful in March, the most hateful tweets in the online conversation in August were more
likely to come from bots. Generally, greater bot activity on a given day further predicted higher levels of hate speech in the entire conversation a week later. Second, bots shifted the targets of hate in online discussion of racism from racial identities to political identities. Whereas organic human-driven hate speech toward Asian and Chinese people was prevalent in March, inorganic bot-fueled hate speech against political actors in the United States became dominant in online discussions of racism in August.

**Bots Amplified Hate in Online Racism Discourse**

Regression analysis revealed that associations between hate speech and bot activity shifted over time ($R^2 = 0.051$, $p < .001$). Figure 1 shows fitted bot scores of accounts depending on the hate speech levels of their tweets. Whereas in March, relationships between hate speech and bot activity were insignificant or negative, consistent positive relationships were observed in August.

In March, the most hateful accounts were more likely to be humans, such that for each standard deviation increase in hate speech score, Tweets Mentioning Racism saw a 0.020 standard deviation drop in bot score ($p < .001$). Discussion-Sparking tweets, however, did not see significant co-variation between bot scores and hate speech scores ($p > .05$). In real terms, a Tweet Mentioning Racism in March with a hate speech score of zero was on average produced by an account with a bot score of 0.592, while a Tweet Mentioning Racism in March with a hate speech score of one was likely to be produced by an account with a bot score of 0.537. This indicates that in the earliest stages of the pandemic, bots and humans alike were precipitating conversations about racism, but hateful talk in response to these prompts largely arose from organic sources.

Conversely, in August, bot-fueled hate speech became particularly prominent. Per standard deviation of increase in hate speech score, the associated bot score of Discussion-Sparking Tweets increased by 0.058 standard deviations ($p < .001$), while Tweets Mentioning Racism saw a similar increase of 0.041 standard deviations in bot score ($p < .001$). In real terms, Discussion-Sparking Tweets in August having a hate speech score of zero were on average produced by an account with a bot score of 0.400, while those having a hate speech score of one were on average produced by an account with a bot score of 0.561. Tweets Mentioning Racism in August having a hate speech score of zero were on average produced by an account with a bot score of 0.592, while those with a hate speech score of one were on average produced by an account with a bot score of 0.670. Hence, 6 months into the pandemic, hate speech was largely being produced by inorganic sources, injected in an automated fashion by social bots.

More than being more hateful, bots were further associated with increased hate speech in the online conversation a week later. Figure 2 shows the results of a path model relating bot activity and hate speech with a 1-week lag. Table 3 further breaks down path model estimates into the direct, indirect, and total effects of the variables on future hate speech levels.

Path analysis indicated that even controlling for current levels of hate speech and current levels of bot activity were positively associated with both future bot activity ($p < .001$) and future levels of hate speech ($p < .05$). Bots were thus
linked to increased hate speech in online discussion of racism directly as well as indirectly by predicting greater bot activity in the future.

Remarkably, these effects are greater than the association between current levels of hate speech and future levels of hate speech, which is not significant ($p > .05$). This indicates that the rise or fall of hate speech does not depend strongly on how much hate speech there currently is but rather on how many bots there are.

In other words, bot activity explained the trajectory of hate speech levels more than hate speech levels itself, highlighting the central role of automated actors in inorganically amplifying hate in online discourse about racism. This effect was robust to whether the tweet was a Tweet Mentioning Racism or a Discussion-Sparking Tweet, and whether it was tweeted in March or August.

**Bots Shift Talk From Racial to Political Identities**

Having established that bots amplified hate in online discussions about racism, we then probed which identities served as the target of that hate. This was crucial as it helped quantify measures beyond how much hate was present in online discourse on racism during the pandemic and inquired into how that hate was directed.

To investigate this question, Figure 3 shows the fitted logistic regression models which estimate the probability that a given tweet mentions different classes of identities given the tweet’s hate speech score. Table 4 specifically quantifies these hate-identity associations by summarizing the average change in the log-odds of identity classes being mentioned given a 0.10 increment in the tweet’s hate speech score.
The results of this analysis indicate that identity associations were generally higher among Discussion-Sparking Tweets compared to Tweets Mentioning Racism. In other words, it was much more likely for Discussion-Sparking Tweets to invoke certain identities in a hateful manner, subsequently prompting Tweets Mentioning Racism. These dynamics could potentially encompass the calling out of racist rhetoric in the initial Discussion-Sparking Tweets, but the results of the path model shown in Table 3 suggest that Discussion-Sparking Tweets were less likely to inspire further hate in the future. Evidence of the hypothesized “call-out” dynamic is therefore nuanced; although hateful Discussion-Sparking Tweets were more likely to invoke specific identities, they may not have been as successful at stimulating further hate as the Tweets Mentioning Racism themselves.

More pointedly, with respect to which identity classes were targeted, it was evident that across both time periods, the strongest hate-identity associations were detected among racial identities and political identities. These associations were particularly important to examine in relation to who mentioned these identities, and when they did so. In March, a tweet with a 0.50 hate speech score had the highest probability—80.74%—of mentioning a racial identity when uttered by a human in a Discussion-Sparking Tweet. But this hate-identity association dropped from 0.781 in March to 0.614 in August. Meanwhile, in August, a tweet with a 0.50 hate speech score had the highest probability—83.92%—of mentioning a political identity when uttered by a bot in a Discussion-Sparking Tweet. This hate-identity association rose from 0.784 in March to 0.890 in August. Hate-identity associations with gender identities and religious identities were less notable across all time and tweet types.

Between March and August, then, we find evidence that (a) hate speech in online discussions of racism shifted from being about racial identities to political identities and that (b) this shift was fueled significantly by bots. Organic, human-driven hate in relation to racial identities thrived in online conversations about racism during the earliest weeks of the COVID-19 outbreak, but inorganic, bot-driven political hate dominated these discussions 6 months into the pandemic. Yet which identities, in particular, constituted these shifting targets? Figure 4 answers this question by quantifying which specific identity terms were most salient across time periods and which ones were disproportionately mentioned by bots. Table 5 further quantifies how these bot proportions changed over time with statistical significance. Masked examples of hate speech associated with racial and political identities are provided in Table 6. These examples show how hate speech targeting racial identities can express sinophobic tropes associating Chinese people with the pandemic as well as how hate speech targeting political groups can dehumanize and threaten both conservative and liberal political figures in the United States.

Early in March, the identities bots used more than expected were people, racist, and doctor for Discussion-Sparking Tweets, but these differences were not statistically significant ($p > .05$). On the other hand, American ($p < .001$), Asian American ($p < .001$), bigot ($p < .001$), scapegoat ($p < .001$), journalist ($p < .001$), and Democrat ($p < .001$) were significantly more salient than expected among bots for Tweets Mentioning Racism. Collectively, these suggested that bots had also been discussing xenophobia against Asian populations. However, even as the broader human conversation was about Asian people in general, bot-driven messages were already discussing racism within the realm of American society and politics. This was evident in references to local political parties and America-specific minorities of Asian descent. Hence, though not as salient as in August, bots in March were already targeting both racial and political discourse specific to the United States.

Later on in August, bots appeared to focus disproportionately on the terms racist ($p < .001$), president ($p < .001$), American ($p < .001$), Democrat ($p < .001$), voter ($p < .001$), traitor ($p < .001$), doctor ($p < .001$), Russian ($p < .001$), mayor ($p < .001$), and bigot ($p < .001$) for Discussion-Sparking Tweets. In Tweets Mentioning Racism, they instead focused on the terms racist ($p < .001$), American ($p < .001$), friend ($p < .001$), president ($p < .001$), leader ($p < .001$), lady ($p < .001$), voter ($p < .001$), Republican ($p < .001$), and thug ($p < .05$). Here, the amplification of political confrontation is immediately apparent, as more pejorative and political terms dominated the top identities disproportionately invoked by bots in their discussions of racism. Moreover, the bot mentions of voters further point to their apparent election-oriented strategic objectives.

Between the two time periods, we detected a large number of statistically significant changes in the proportion of bots mentioning key identity terms. For Tweets Mentioning Racism, there were notable decreases in the proportion of bots containing the racial identities such as Asian ($p < .05$) and Chinese ($p < .001$), and notable increases in the proportion of bots mentioning the political identities American ($p < .001$) and president ($p < .001$). Although there was a concurrent decrease for the term Democrat ($p < .001$), we note that its raw bot proportion remained high relative to other identity terms. In fact, the term already had a higher than expected bot proportion in March, so even with a decrease in August, the term Democrat remained a salient identity in bot-fueled discourse.

Meanwhile, for Discussion-Sparking Tweets, we saw increases across the board in bot proportions for all identities. This means that between March and August, bots became much more active in mentioning all kinds of identity terms, including both racial and political terms. This may speak to the intensification of information operations around pandemic discussions about racism such that the targeting of identities became a prominent tactic. But consistent with our previous analysis, whereas the bot proportions for racial identities like Asian and Chinese saw three- to five-fold increases ($p < .001$), the political identity of president in particular saw a ten-fold increase ($p < .001$).
Discussion

Online discourse surrounding COVID-19 has featured large-scale discussions of racism, reflecting broader issues of racial disparities during the pandemic (Chiriboga et al., 2020; Khazanchi et al., 2020; Reny & Barreto, 2022). As our findings demonstrate, however, the locus of online discourse on racism was neither fixed nor free from inorganic manipulation. Rather, it was subject to the influence of bot-driven information operations which, over time, (a) amplified levels of hate speech and (b) redirected its focus from targeting Asian groups to discussing politics related to the 2020 US elections.

These findings meaningfully extend scholarship at the intersections of hate speech, social bots, racism, and the COVID-19 pandemic. Here, we have highlighted the multifaceted linkages between these phenomena. Literature on pandemic-fueled hate speech has thus far predominantly focused on anti-Asian and anti-Chinese sentiments in the early months of the pandemic (Dubey, 2020; Stechemesser et al., 2020; Uyheng & Carley, 2021a), and online conversations about racism have largely been assumed to reflect organic xenophobia or community resilience to racial violence (Abd-Alrazaq et al., 2020; Abidin & Zeng, 2020; Kuo et al., 2020; Tsao et al., 2021). Our work underscores that in moments of crisis, as such social divides come to the fore (Clissold et al., 2020; Elias et al., 2021), their discursive potency may become ripe targets for information campaigns seeking to exploit salient societal tensions as strategic vulnerabilities (Arif et al., 2018; Starbird, 2019). Although this is often exploited and manipulated by people organically perpetuating hate speech on social media, our results show how information operations can inorganically influence online hate speech in an automated manner. From this standpoint, the online world acts not only as a mirror of offline contentions but also as a space in which they can be exacerbated or transformed, with meaningful
public health consequences (Awan & Zempi, 2016; Hoover et al., 2021; Weber et al., 2020).

Meanwhile, studies on bot activity during the COVID-19 outbreak have dealt primarily with low-credibility information and conspiracy theories (Ferrara, 2020; Moffitt et al., 2021; Xu & Sasahara, 2022), thus omitting their central role in sparking conflicts (Kim & Kesari, 2021; Uyheng & Carley, 2020). In specifically foregrounding how bots drove hateful political talk around racism in August (particularly in relation to US politics), we have showed precisely that the activities of bots are also organized around amplifying contentions and not just spreading falsehoods. Furthermore, while evidence is mixed regarding the impacts of bot activities during the pandemic (Himelein-Wachowiak et al., 2021; Xu & Sasahara, 2022), we found that in the online conversation on racism, bots' greater salience triggered higher levels of hate speech over time. Contributing to the rich scholarship on bot-driven information operations during high-profile

### Table 5. Results of Two-Sample Proportion Tests Across Identities Mentioned by Bots in March and August 2020.

| Dataset | Identity | Change | Bot % (Mar.) | Bot % (Aug.) | $\chi^2$ | $p$  |
|---------|----------|--------|--------------|--------------|---------|------|
| Tweets Mentioning Racism | racist | Increase | 38.04 | 40.91 | 136.05 | <.001*** |
| | Asian | Decrease | 35.63 | 33.78 | 6.05 | .0139* |
| | people | Increase | 32.16 | 32.73 | 1.29 | .2561 |
| | Chinese | Decrease | 37.35 | 30.03 | 33.32 | <.001*** |
| | American | Increase | 41.32 | 43.73 | 12.47 | <.001*** |
| | president | Increase | 38.41 | 48.27 | 128.35 | <.001*** |
| | man | Increase | 34.94 | 37.8 | 5.23 | 0.0222* |
| | bigot | Decrease | 46.64 | 40.23 | 17.27 | <.001*** |
| | democrat | Decrease | 49.65 | 39.66 | 32.51 | <.001*** |
| | woman | Decrease | 40.64 | 36.15 | 8.26 | .0041** |

**Discussion-Sparking Tweets**

| Identity | Change | Bot % (Mar.) | Bot % (Aug.) | $\chi^2$ | $p$ |
|----------|--------|--------------|--------------|---------|------|
| Chinese | Increase | 3.87 | 19.31 | 129.35 | <.001*** |
| people | Increase | 8.6 | 18.92 | 58.92 | <.001*** |
| racist | Increase | 8.51 | 33.48 | 223.15 | <.001*** |
| official | Increase | 2.17 | 3.95 | 0.26 | .6109 |
| American | Increase | 3.3 | 45.8 | 307.79 | <.001*** |
| Asian | Increase | 4.82 | 14.92 | 21.56 | <.001*** |
| president | Increase | 3.98 | 33.17 | 161.33 | <.001*** |
| man | Increase | 5.14 | 17.71 | 15.79 | <.001*** |
| democrat | Increase | 4.4 | 24.64 | 32.68 | <.001*** |
| republican | Increase | 5.26 | 11.55 | 4.62 | .0315* |
| woman | Increase | 6.25 | 13.36 | 3.22 | .0726 |
| someone | Increase | 7.63 | 11.69 | 0.82 | .3652 |
| doctor | Increase | 11.54 | 54.38 | 59.89 | <.001*** |

Note: *$p<.05$, **$p<.01$, ***$p<.001$.

### Table 6. Samples of Hate Speech Tweets Targeting Racial and Political Identities.

| Tweet type | Racial hate | Political hate |
|------------|-------------|----------------|
| Discussion-Sparking Tweet | @[Mention] Son of b[*****] is Chinese. . . f[****] the Chinese people! I proudly say that I am racist against the Chinese! Do you eat bat, m[**********]? I made bat soup for you, would you eat, b[******]? | @[Mention] God damm, you are a massive piece of f[****]. A orange spray-tanned, rug-wearing, thieving, incompetence, racist, xenophobic piece of f[****]. I hope YOU get the Coronavirus, and then you'll feel the pain of all the people that have died and the suffering their families dealt with. |
| Tweet Mentioning Racism | Yup! CHINA IS THE LAND OF ALL VIRUSES. SORRY NOT SORRY. NOT BEING RACIST BUT f[***] ALL CHINESE PEOPLE. | @[Mention] @[Mention] OH f[*****]! THE CORRUPT DEMS, HOLLYWOOD f[*****] boys (& L[****] Ellen) R SO, SO F[******]! ALL THE SLEAZY LIB f[****] CAN DO IS KEEPING PUSHING ALONG WITH THEIR BLM HATE GROUP & THEIR FAKE RACISM NARRATIVES, & ATTEMPTING TO BRAINWASH ALL AMERICANS ABOUT COVID-19 KILLING US ALL! TRUMP 2020 |

Identifiable information and abusive terms are masked.
events like electoral contests for power (Ferrara et al., 2020; King et al., 2020; Woolley & Howard, 2016), evidence from this study thus illustrates how these activities do not abate during as disruptive a crisis like the COVID-19 pandemic but rather strategically assimilate their concerns in fueling emergent conflicts.

Methodologically, this work points to several insights for the study of online hate speech over large-scale online conversations (Bluic et al., 2018; Chetty & Alathur, 2018). While state-of-the-art efforts at classifying hate speech as text are valuable (Davidson et al., 2017; Fortuna & Nunes, 2018), we have presented a framework for more sensitively characterizing how hate speech is manipulated, and to what tactical ends (Starbird, 2019). In particular, as argued by Reddi et al. (2021) and Kuo and Marwick (2021), it matters which social divides are exploited, such as the racially charged messaging in March versus the political conflict in August, and how they become linked to one another. Our integration of identities as a conceptual anchor into understanding a problem-based phenomenon exemplifies the fusion of social scientific theory and computational methods as advocated in the growing field of social cybersecurity (K. M. Carley, 2020; Uyheng et al., 2020).

From the onset of the COVID-19 pandemic in 2020, its racialized facets have been central to the ways people have understood, experienced, and responded to the global crisis (Gee et al., 2020; Tan et al., 2022). Targeting calls for enhanced resilience to public health misinformation and rectifying societal disparities, this research stresses the need to attend to the ways online manipulation specifically mobilizes identities to sow discord and perpetuate harms (Freelon & Wells, 2020). Recognizing their interlinked nature means that paths forward similarly entail coordinated efforts which counter information campaigns in a socially sensitive manner while social divides are in turn addressed with robust protection from malicious coordinated influence.

Conclusion

This article examined Twitter discussions of racism during the early months of the COVID-19 pandemic. By combining social scientific theory about identities with computational tools for hate speech and social bot detection, we empirically characterized bot-fueled amplification and redirection of hate from focusing on Asian and Chinese populations in March to targeting political discourse surrounding US political figures in August. This work advances research on how hate speech targets particular social groups and, in turn, how bot-driven information operations can take advantage of societal divides to achieve strategic objectives during times of crisis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the Knight Foundation and the Office of Naval Research (ONR) Grants N000141812106 and N000141812108. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) and the Center for Informed Democracy and Social Cybersecurity (IDeAS). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, ONR, or the US government.

ORCID iD

Joshua Uyheng 12 https://orcid.org/0000-0002-1631-6566

References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study. Journal of Medical Internet Research, 22(4), Article e19016.

Abidin, C., & Zeng, J. (2020). Feeling Asian together: Coping with #covidracism on Subtle Asian Traits. Social Media+ Society, 6(3), 1–5.

Altman, N., Carley, K. M., & Reminga, J. (2017). ORA user’s guide 2017 (Technical Report). CASOS Center, Institute for Software Research, Carnegie Mellon University.

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMatter discourse. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), Article 20.

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. Aggression and Violent Behavior, 27, 1–8.

Beskow, D. (2020). Finding and characterizing information warfare campaigns [PhD Thesis, Carnegie Mellon University].

Beskow, D. M., & Carley, K. M. (2018). Bot-hunter: A tiered approach to detecting & characterizing automated activity on Twitter. In R. Thomson, C. Dancy, A. Hyder & H. Bisgin (Eds.), International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (Vol. 3, p. 3). Springer.

Beskow, D. M., & Carley, K. M. (2019). Social cybersecurity: An emerging national security requirement. Military Review, 99(2), 117.

Bluic, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. Computers in Human Behavior, 87, 75–86.

Buerger, C. (2021). #iamhere: Collective counterspeech and the quest to improve online discourse. Social Media+ Society, 7(4), 1–17.

Byrne, B. M. (2005). Factor analysis: Confirmatory. In: B. Everitt & D. Howell (Eds.), Encyclopedia of statistics in behavioral science. Wiley. https://doi.org/10.1002/0470013192.bsa130

Carley, K. M. (2014). ORA: A toolkit for dynamic network analysis and visualization. In R. Alhajj & J. Rokne (Eds.), Encyclopedia of social network analysis and mining (pp. 1219–1228). Springer
Carley, K. M. (2020). Social cybersecurity: An emerging science. *Computational and Mathematical Organization Theory, 26*(4), 365–381.

Carley, L. R., Reminga, J., & Carley, K. M. (2018). ORA & NetMapper. In R. Thomson, C. Dancy, A. Hyder & H. Bisgin (Eds.), International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer.

Chan, C., & Montt Strabucchi, M. (2021). Many-faced orientalism: Racism and xenophobia in a time of the novel coronavirus in Chile. *Asian Ethnicity, 22*(2), 374–394.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior, 40*, 108–118.

Chiriboga, D., Garay, J., Buss, P., Madrigal, R. S., & Rispel, L. C. (2020). Health inequity during the COVID-19 pandemic: A cry for ethical global leadership. *The Lancet, 395*(10238), 1690–1691.

Clissold, E., Nylander, D., Watson, C., & Ventriglio, A. (2020). Pandemics and prejudice. *International Journal of Social Psychiatry, 66*(5), 421–423.

Daniels, C., DiMaggio, P., Mora, G. C., & Shepherd, H. (2021). Has pandemic threat stoked xenophobia? How COVID-19 influences California voters’ attitudes toward diversity and immigration. *Sociological Forum, 36*, 889–915.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language [Conference session]. Eleventh International AAAI Conference on Web and Social Media, Montreal, Canada.

Devakumar, D., Shannon, G., Bhopal, S. S., & Abubakar, I. (2020). Racism and discrimination in COVID-19 responses. *The Lancet, 395*(10231), 1194.

Dhanani, L. Y., & Franz, B. (2021). Why public health framing matters: An experimental study of the effects of COVID-19 framing on prejudice and xenophobia in the United States. *Social Science & Medicine, 269*, Article 113572.

Dubey, A. D. (2020). The resurgence of cyber racism during the COVID-19 pandemic and its aftereffects: Analysis of sentiments and emotions in tweets. *JMIR Public Health and Surveillance, 6*(4), Article e19833.

Egede, L. E., & Walker, R. J. (2020). Structural racism, social risk factors, and Covid—19—A dangerous convergence for Black Americans. *New England Journal of Medicine, 383*(12), Article e77.

Elías, A., Ben, J., Mansouri, F., & Paradies, Y. (2021). Racism and nationalism during and beyond the COVID-19 pandemic. *Ethnic and Racial Studies, 44*(5), 783–793.

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*. https://firstmonday.org/article/view/10633/9548

Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 US presidential election. *First Monday*. https://firstmonday.org/ojs/index.php/fm/article/view/11431

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96–104.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1–30.

Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication, 37*(2), 145–156.

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the risks of “infodemics” in response to COVID-19 epidemics. *Nature Human Behaviour, 4*, 1285–1293. https://doi.org/10.1038/s41562-020-00994-6

Gee, G. C., Ro, M. J., & Rimoin, A. W. (2020). Seven reasons to care about racism and COVID-19 and seven things to do to stop it. *American Journal of Public Health, 110*(7), 954–955.

Gover, A. R., Harper, S. B., & Langton, L. (2020). Anti-Asian hate crime during the COVID-19 pandemic: Exploring the reproduction of inequality. *American Journal of Criminal Justice, 45*(4), 647–667.

Heise, D. R. (1987). Affect control theory: Concepts and model. *Journal of Mathematical Sociology, 13*(1–2), 1–33.

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., Epstein, D. H., Leggio, L., & Curtis, B. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research, 23*(5), Article e26933.

Hoover, J., Atari, M., Mostafazadeh Davani, A., Kennedy, B., Portillo-Wightman, G., Yeh, L., & Dehghani, M. (2021). Investigating the role of group-based morality in extreme behavioral expressions of prejudice. *Nature Communications, 12*(1), 1–13.

Huang, B., & Carley, K. M. (2020). Disinformation and misinformation on Twitter during the novel coronavirus outbreak. arXiv preprint arXiv:2006.04278.

Islam, M. S., Sarkar, T., Khan, S. H., Kamal, A. H. M., Hasan, S. M., Kabir, A., Yeasmin, D., Islam, M. A., Chowdhury, K. I. A., Anwar, K. S., Chughtai, A. A., & Seale, H. (2020). COVID-19-related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene, 103*(4), 1621–1629.

Joseph, K., Wei, W., Benigni, M., & Carley, K. M. (2016). A social-event based approach to sentiment analysis of identities and behaviors in text. *The Journal of Mathematical Sociology, 40*(3), 137–166.

Khazanchi, R., Evans, C. T., & Marcelin, J. R. (2020). Racism, not race, drives inequity across the COVID-19 continuum. *JAMA* *Network Open*, 3(9), Article e2019933.

Kim, J. Y., & Kesari, A. (2021). Misinformation and hate speech: The case of Anti-Asian hate speech during the COVID-19 pandemic. *Journal of Online Trust and Safety*, 1(1). https://tsjournal.org/index.php/jots/article/view/13

King, C., Bellutta, D., & Carley, K. M. (2020). Lying about lying on social media: A case study of the 2019 Canadian elections. In R. Thomson, H. Bisgin, C. Dancy, A. Hyder & M. Hussain (Eds.), *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 75–85). Springer.

Kuo, R., & Marwick, A. (2021). Critical disinformation studies: History, power, and politics. *Harvard Kennedy School Misinformation Review, 2*(4), 1–11.

Kuo, R., Zhang, A., Shaw, V., & Wang, C. (2020). #FeministAntibodies: Asian American media in the time of coronavirus. *Social Media+ Society, 6*(4), 1–11.

Le, T. K., Cha, L., Han, H. R., & Tseng, W. (2020). Anti-Asian xenophobia and Asian American COVID-19 disparities. *American Journal of Public Health, 110*(9), 1371–1373.
Lozada, F. T., Seaton, E. K., Williams, C. D., & Tynes, B. M. (2021). Exploration of bidirectionality in African American and Latinx adolescents’ offline and online ethnic-racial discrimination. *Cultural Diversity and Ethnic Minority Psychology, 27*, 386–396.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE, 14*(8), Article e0221152.

Moffitt, J., King, C., & Carley, K. M. (2021). Hunting conspiracy theories during the COVID-19 pandemic. *Social Media + Society, 7*(3), 1–17.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013, July). *Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter’s firehose* [Conference session]. Proceedings of the International AAAI Conference on Web and Social Media, Boston, MA, USA.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.

Poljomen, M., & Udupa, S. (2017). Extreme speech online: An anthropological critique of hate speech debates. *International Journal of Communication, 11*, 1173–1189.

Reddi, M., Kuo, R., & Kreiss, D. (2021). Identity propaganda: Racial narratives and disinformation. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448211029293

Reichelmann, A., Hawdon, J., Costello, M., Ryan, J., Blaya, C., Llorent, V., Oksanen, A., Räsänen, P., & Zych, I. (2021). Hate knows no boundaries: Online hate in six nations. *Deviant Behavior, 42*(9), 1100–1111.

Reny, T. T., & Barreto, M. A. (2022). Xenophobia in the time of pandemic: Othering, anti-Asian attitudes, and COVID-19. *Politics, Groups, and Identities, 10*, 209–232.

Starbird, K. (2019). Disinformation’s spread: Bots, trolls and all of us. *Nature, 571*(7766), 449–450.

Stechemesser, A., Wenz, L., & Levermann, A. (2020). Corona crisis fuels racially profiled hate in social media networks. *EClinicalMedicine, 23*, Article 100372. https://doi.org/10.1016/j.eclinm.2020.100372

Tan, S. B., DeSouza, P., & Raifman, M. (2022). Structural racism and COVID-19 in the USA: A county-level empirical analysis. *Journal of Racial and Ethnic Health Disparities, 9*, 236–246.

Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: A scoping review. *The Lancet Digital Health, 3*, Article e175–e194.

Uyheng, J., & Carley, K. M. (2020). Bots and online hate during the COVID-19 pandemic: Case studies in the United States and the Philippines. *Journal of Computational Social Science, 3*, 445–468.

Uyheng, J., & Carley, K. M. (2021a). Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Applied Network Science, 6*(1), 1–21.

Uyheng, J., & Carley, K. M. (2021b). An identity-based framework for generalizable hate speech detection. In R. Thomson, C. Dancy, A. Hyde & H. Bisgin (Eds.), International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 121–130). Springer.

Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., & Carley, K. M. (2020). Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. *Computational and Mathematical Organization Theory, 26*, 465–483.

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Alexander Haslam, S., Jetten, J., . . . Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour, 4*, 460–471.

Wachs, S., Bilz, L., Wettstein, A., Wright, M. F., Krause, N., Ballaschl, C., & Kansok-Dusche, J. (2022). The online hate speech cycle of violence: Moderating effects of moral disengagement and empathy in the victim-to-perpetrator relationship. *Cyberpsychology, Behavior, and Social Networking, 25*, 223–229.

Weber, M., Viehmann, C., Ziegele, M., & Scheler, C. (2020). Online hate does not stay online: How implicit and explicit attitudes mediate the effect of civil negativity and hate in user comments on prosocial behavior. *Computers in Human Behavior, 104*, Article 106192.

Woollery, S. C., & Howard, P. N. (2016). Political communication, computational propaganda, and autonomous agents: Introduction. *International Journal of Communication, 10*, 4882–4890.

Xu, W., & Sasahara, K. (2022). Characterizing the roles of bots on Twitter during the COVID-19 infodemic. *Journal of Computational Social Science, 5*, 591–609.

**Author Biographies**

Joshua Uyheng (MS Carnegie Mellon University) is a societal computing PhD student at Carnegie Mellon University, where he is advised by Dr. Kathleen M. Carley at the Center for Computational Analysis of Social and Organizational Systems (CASOS), and a Knight Fellow at the Center for Informed Democracy and Social Cybersecurity (IDeaS). His research examines the narrative and network dynamics of online hate, focusing especially on the Asia-Pacific and the Global South.

Daniele Bellutta (MSc Dartmouth College) is a societal computing PhD student for Software Research at Carnegie Mellon University. His research interests lie at the intersection of computer science and political science.

Kathleen M. Carley (PhD Harvard, H.D. University of Zurich) is Professor of Computer Science at the Institute for Software Research at Carnegie Mellon University, IEEE Fellow, Director of the Center for Computational Analysis of Social and Organizational Systems (CASOS), Director of the Center for Informed Democracy and Social Cybersecurity (IDeaS), and CEO of Netanomics. She is the recipient of the USGA Academic Award at GEONT 2018 for her work on geospatially enabled dynamic network analytics, the Allen Newell award for research excellence, the Lifetime Achievement Award from the Sociology and Computers Section of the ASA (2001), and the Simmel Award for advances in social networks from INSNA (2011). Her research combines cognitive science, sociology, and computer science to address complex social and organizational issues. Her pioneering research led to the areas of computational social science, dynamic network analysis, and social cybersecurity. She has over 400 publications and has served on multiple National Academies panels.