From Xenophobia to Political Confrontation: Shifting Online Discussions of Racism During the COVID-19 Pandemic [*]

Joshua Uyheng[1][0000−0002−1631−6566], Daniele Bellutta[1][0000−0002−6131−9846], and Kathleen M. Carley[1][0000−0002−6356−0238]

CASOS Center, Institute for Software Research, Carnegie Mellon University,
Pittsburgh PA 15213, USA
{juyheng, dbellutt, kathleen.carley}@cs.cmu.edu

Abstract. The ongoing COVID-19 pandemic has stoked racism and racial division worldwide. Many early studies indicated that xenophobic hate toward Asian populations had taken root especially in online discourse. However, as social and political conditions have evolved with time, the nature of hate speech and racism have likewise shifted in the public conversation. Using a large-scale, long-term dataset of tweets about racism during the pandemic, we analyze differences in the prevalence and targeting of hate speech. We empirically demonstrate that while hateful racism discourse primarily featured "Asian" and "Chinese" identities in March, focus was redirected towards "American" and the "President" in August. Collectively, these findings suggest that online discussions of racism shifted from xenophobia at the onset of the pandemic to critical attacks against political leaders like the U.S. president. Our findings bear implications for understanding the political potency of racism during the pandemic, as well as the computational study of online hate speech more broadly.

Keywords: Hate speech · Racism · COVID-19 · Identities

1 Introduction

Over the past year, the COVID-19 pandemic has not only triggered public health crises worldwide but also exacerbated social conflicts [3, 16]. Burgeoning research points to the seriousness of online hate speech around COVID-19, particularly in relation to xenophobic and racist discourse. During the nascent stages of the

[*] This work was supported in part by the Knight Foundation and the Office of Naval Research 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 in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the U.S. government.

pandemic, racial, ethnic, and national groups – especially Asian and Chinese populations – became associated with the pathogen through negative political and media representations [9, 12]. Moreover, recent work points to the bot-driven amplification of hateful, racialized content in relation to the pandemic, with potential consequences for sowing discord in ethnically diverse societies struggling with local outbreaks [14]. Racism thus plays a crucial role in aggravating the pandemic's diverse and unequal impacts, and occupies a key position in public discourse surrounding the disease.

Months later, however, the social dimensions of the pandemic have transformed. In this view, we posit that public understandings of the disease – as well as its attendant associations with various social groups – have similarly evolved. However, extant research tends to be focused on earlier periods of the pandemic. For instance, to our knowledge, no existing work quantifies temporal shifts in online racist discourse around the pandemic. From an identity perspective, it also remains to be seen whether racism remains particularly sinophobic or has become repurposed for other targets. Using a series of interoperable computational tools [15], this paper examines these changes empirically through the lens of social cybersecurity [1]. Nearly a year into the pandemic, this work aims to contribute to longer-term scholarship on racism and the COVID-19 pandemic while also introducing a straightforward quantitative methodology for characterizing changes in hate speech over time.

In sum, we ask the following research questions:

1. How much hate speech is associated with discussions of racism during the COVID-19 pandemic?
2. What are the identities associated with racism and hate speech in online conversations about the pandemic?
3. How do patterns of hate speech around the online discussion of racism vary between early on in the pandemic and several months later?

2 Methods

2.1 Dataset

We examined online COVID-19 discussions using a large-scale global dataset collected by Huang and Carley [7]. The full dataset consisted of over 200 million pandemic-related tweets obtained using the Twitter streaming API with search terms related to COVID-19.

We were particularly interested in the prevalence of hate speech in the Twitter coronavirus conversation at two points in time – March 2020 (Time 1) and August 2020 (Time 2). Hence, over this time period, we filtered for tweets that mentioned words from a multilingual list of racism-related terms. Table 1 lists these filter words, such as "racist", "bigot", and "xenophobic". Our hypothesis was that these tweets would likely have been sent in response to instances of racism on Twitter. One meaningful interaction between such tweets, for instance, would entail calling out other users as being racist.

Table 1. Racism-related terms used to filter the data, along with the number of languages other than English into which they were translated.

Terms	Trans.	Terms	Trans.	Terms	Trans.
racism	76	discriminative	58	bigoted	57
racist	72	discrimination	77	xenophobia	68
racial	71	discriminated	74	xenophobe	35
discriminate	73	bigotry	69	xenophobes	36
discriminates	67	bigot	50	xenophobic	60
discriminatory	74	bigots	42		

Table 2. Summary of Twitter datasets on racism around COVID-19.

Dataset	Time 1 (March)		Time 2 (August)	
	Tweets	Users	Tweets	Users
Racism Tweets	518K	425K	176K	139K
Replied Tweets	28K	14K	19K	13K

To investigate these dynamics further, we filtered our tweets mentioning racism to find replies to other tweets. We then collected as many of the tweets that had been replied to as was possible at the time (since some of the original tweets may have been deleted). According to our hypothesis, these tweets that had gathered replies mentioning racism should have been examples of racist activity on Twitter. We therefore ended up with two datasets per point in time: a set of tweets mentioning racism and a set of tweets to which people had replied with mentions of racism. We have summarized statistics on these datasets in Table 2.

2.2 Hate Speech Detection and Characterization

In this study, we detected hate speech by using a machine learning model trained on a seminal benchmark dataset of hate speech [4]. Using the NetMapper software [2], we obtain a feature set of lexical counts derived from the psycholinguistic literature [10, 13]. The model employs a random forest classifier to predict hate speech labels using these features. Prior research has shown that the model achieves a weighted F1 score of 83% on the benchmark data [14]. A tweet was counted as hateful if the model gave it a score above 50%. The proportion of hateful tweets was computed for each dataset, and two-sample proportion tests for equality (with continuity correction) were performed to compare the datasets over time. Proportion tests were done to compare the tweets mentioning racism to the tweets to which they had replied.

Since a key question on the coronavirus racism conversation dealt with which groups were being targeted, we further counted the instances of the multilingual identity terms used by NetMapper [2, 8]. These identity terms were manually

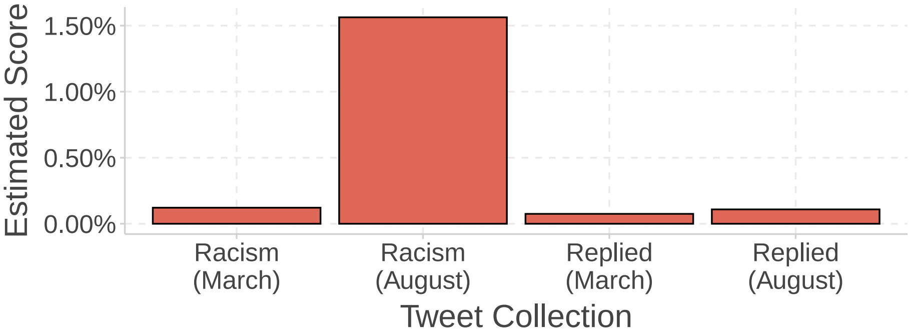

Fig. 1. Predicted proportions of hate speech in racism conversation around COVID-19.

Table 3. Results of proportion tests comparing amounts of hate speech across datasets.

Group 1	Group 2	χ^2	p
Racism Time 1	Racism Time 2	5634.2	$< .001^{***}$
Replied Time 1	Replied Time 2	1.0314	0.310
Racism Time 1	Replied Time 1	2.6738	0.102
Racism Time 2	Replied Time 2	459.32	$< .001^{***}$

coded by two coders as belonging to five categories: gender, politics, race, religion, and other. During the coding process, all identities were allowed to belong to more than one of the first four categories. If an identity did not belong to any of the first four, then it was coded as “other”. Annotations were initially performed in an independent fashion, then subsequently resolved by consensus between the two coders.

Using this identity lexicon, we counted the number of times each term was used in our datasets of tweets and aggregated the counts across the five categories of identity terms. In this scheme, we assumed that a tweet using an identity term that was also classified as hate speech was expressing hate towards the corresponding identity term category.

3 Results

3.1 Hate Speech in COVID-19 Racism Discussion

Figure 1 shows the proportion of hateful tweets for each of the four data sets. Interestingly, across datasets, we find that the overall levels of hate speech in the discussion of racism around COVID-19 are relatively low. Using our hate speech model, most values are below 1%, with the highest detected proportion being around 1.56% for tweets explicitly mentioning racism in August.

Using a series of two-sample tests for equality of proportion, we determined the statistical significance of the differences in the prevalence of hate speech

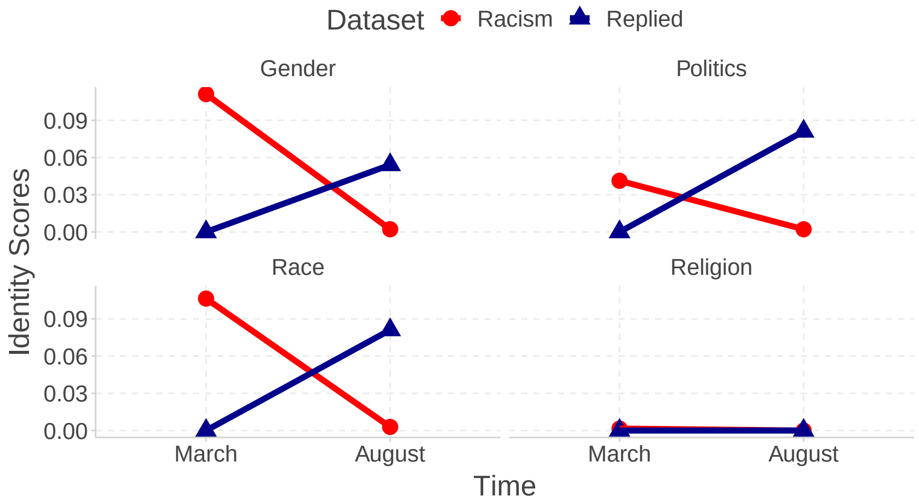

Fig. 2. Changes in average levels of identity mentions in tweets classified as hate speech using a machine learning model.

between our datasets. These results are summarized in Table 3. Based on the hate speech model, we find that the amount of hate appears to significantly increase from March to August among tweets explicitly mentioning racism ($\chi^2 = 5634.2, p < .001$).

During that later time period, the amount of hate is also significantly higher among tweets mentioning racism relative to the tweets to which they are replying ($\chi^2 = 459.32, p < .001$). This is in contrast to the relationship in March, when tweets mentioning racism do not differ significantly in hateful content compared to the tweets they reply to ($\chi^2 = 2.6738, p = 0.102$).

Collectively, the results of this analysis suggest that there were more hateful tweets among August tweets mentioning racism rather than the tweets to which they were replying. This goes against our call-out hypothesis. Instead, the mentions of racism themselves appear to be used in a pejorative fashion.

3.2 Pivoting from Racial to Political Hate

We nuance our analysis of hate speech around the racism discussion by measuring mentions of particular classes of identity terms. Figure 2 shows changes in the rate at which different identities are mentioned in tweets classified as hate speech by our model across datasets. Table 4 further reports the results of two-way analysis of variance (ANOVA) tests which ascertain the statistical significance of differences over time and between racism tweets compared to the tweets to which they reply.

Quite consistently, we see a reversal in the relative level at which identity terms are mentioned among tweets mentioning racism versus those to which

Table 4. Results of two-way analysis of variance (ANOVA) tests comparing identity scores relative to time, dataset type, and their interaction effect. Asterisks indicate levels of statistical significance.

Identities	Dataset		Time		Interaction	
	F	p	F	p	F	p
Gender	0.201	0.654	284.526	$< .001^{***}$	12.037	$< .001^{***}$
Politics	7.050	$< .01^{**}$	60.110	$< .001^{***}$	9.778	$< .01^{**}$
Race	2.077	0.150	253.170	$< .001^{***}$	15.286	$< .001^{***}$
Religion	0.029	0.8659	4.342	$< .05^{*}$	0.082	0.775

they reply. Initially, tweets mentioning racism also mention various identity terms at a higher level than their reply targets. But in August, it is the reply targets which are invoking identity terms at greater rates. These crossovers are captured by statistically significant interaction effects for gender identities ($F = 12.037, p < .001$), political identities ($F = 9.778, p < .01$), and racial identities ($F = 15.286, p < .001$). We additionally note that religious identities are scarcely mentioned in our datasets.

Most striking are the changes in racial and political identities. Between March and August, the sharp decline in the association between racism tweets and racial identities signals that hateful tweets mentioning racism are no longer explicitly attacking particular racial groups. By contrast, the increased scores among reply targets suggest that it is these tweets expressing racially charged hate. In other words, later on in the pandemic, we measure behaviors more in line with our previous call-out hypothesis: when netizens tweet hatefully in relation to specific racial identities, other netizens respond to them aggressively with call-outs of racism. These interactions were not as salient in March.

Meanwhile, for political identities, we observe a similar crossover effect. But while the differences were not as stark between datasets in March, the gap widens significantly by August. Notable here is that the level of political identities among reply targets rises to a level similar to that of the racial identities, indicating that the discussion of racism not only concerns racial groups, but also political actors. More specifically, in August, the pattern appears to be that netizens tweet hatefully in relation to key political figures, and netizens likewise respond hatefully with charges of racism.

3.3 From Xenophobia to Political Fallout

For more in-depth analysis of this shift in the meaning of racism, we turn to the specific identities mentioned in the tweets collected. Figure 3 depicts the top twenty identity terms invoked in all four datasets, alongside their prevalence relative to the most frequently occurring identity term in each dataset.

Examination of the terms in March points to the initial dominance of Asian and Chinese identities as topics discussed with racism during COVID-19. For

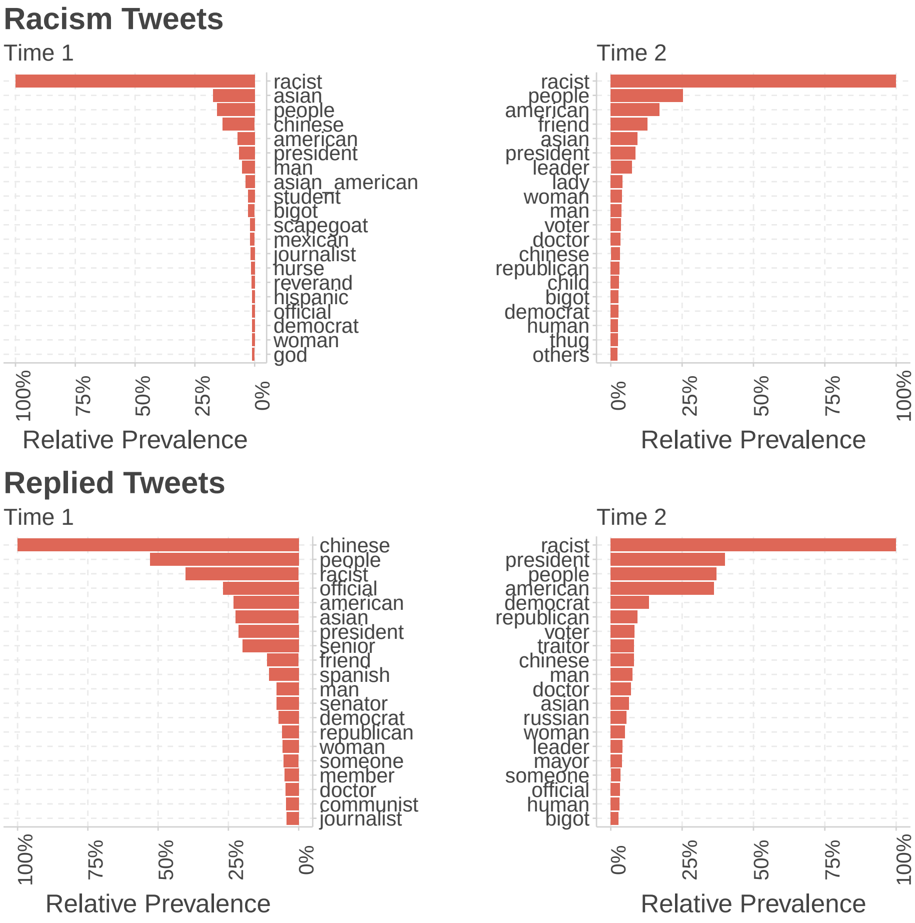

Fig. 3. Relative prevalence of top 20 identities mentioned in online racism discussion. Values are normalized relative to the occurrences of the most frequently used term.

tweets mentioning racism, "Asian" and "Chinese" are the second and fourth ranked identity terms following mentions of racists, with relative prevalence scores of 17.38% and 13.32%. Among reply targets, "Asian" falls to the sixth rank, though with a higher relative prevalence of 22.40%; but "Chinese" is the most mentioned identity term in the entire dataset. These measurements strongly suggest that discussions of racism around COVID-19 responded strongly to discussions of Chinese and Asian groups.

However, by August, the word "Asian" ranks fifth in tweets mentioning racism, with a relative prevalence of 9.35%. During the same time period, "Chinese" ranks thirteenth with a relative prevalence of 3.14%. Among reply targets, "Chinese" falls from the top identity to the ninth with a relative prevalence of 8.15%, while "Asian" occupies the twelfth rank with a relative prevalence of

6.36%. Taken together, these measurements deepen our assessment of reduced discussion around racial groups in August, especially relative to the pervasive targeting of Asian and Chinese populations in March.

At the same time, “American” and “President” rank highly among racism tweets at fifth and sixth spots in March. But, their relative prevalence is initially low at 7.20% and 6.62%, respectively. Among reply targets, their ranks are comparable at fifth and seventh, but they have relatively higher prevalence at 23.17% and 21.46%.

However, by August, tweets about racism rank mentions of “American” at third, with a relative prevalence of 17.03%, and “President” at fifth, with a relative prevalence of 8.66%. Reply targets further feature this increase in rank at a greater rate, with “American” rising to fourth at a relative prevalence of 36.22% and “President” rising to second at a relative prevalence of 39.99%, right after “racist”. Collectively, these measurements point to an intensified focus on U.S. politics, particularly in relation to the nearing U.S. presidential elections in November. In this electoral context, the expression of hate in relation to racism further appears to be presented as criticism of President Trump against the backdrop of the pandemic.

This latter assessment of election-linked criticism is further borne out by other terms among the top twenty identities. By August, mentions of “Democrat” and “Republican” rise substantially among reply targets, ranking fifth and sixth, respectively, indicating their focal role in triggering discussions around racism in relation to the pandemic. Mentions of voters likewise appear among tweets mentioning racism as well as their target replies, ranking eleventh and sixth, respectively. This potentially suggests that racism may be an issue of significance to voters’ deliberative process. Finally, more pejorative political terms likewise appear in the top twenty terms. “Traitor” is the most notable for reply targets, ranking eighth (right above “Chinese”).

4 Limitations

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. Furthermore, our methods may have underestimated the prevalence of hate speech due to limits in the hate speech model’s generalizability and predictive performance. While improvements to hate speech detection methods may certainly be explored using more advanced models [6], our focus in the present work lies primarily with scalability and interpretability. In addition, hate speech may be under-represented in our sets of tweets that were replied to with mentions of racism. Though the tweets discussing racism were collected via a Twitter data stream, the tweets being replied to were necessarily collected after the fact. This means that some tweets could have been removed before we were able to collect them. For March 2020, our data collection was able to collect 84.60% of the tweets with replies mentioning racism. For August 2020, that proportion was 81.88%. It is therefore clear that some tweets were deleted before being collected. However, not all of the inaccessible tweets had been deleted; many were simply set as private. This means that though our datasets of replied tweets may under-represent the hate speech present in the true data, the effect may be small. More research into this issue is necessary before a meaningful conclusion can be made.

5 Conclusions and Future Work

Online discourse around COVID-19 has featured significant discussion of racism, reflecting broader issues of racial inequities around the pandemic [3, 5]. Over time, however, our findings demonstrate that while racism indeed reflected xenophobic hate speech in the early months of the pandemic, it has also become a strikingly political category, anchored now on the actions of political leaders, reflecting their track record of themselves stoking xenophobia or racial divisions in response to the pandemic [11]. These findings meaningfully extend the literature on pandemic-fueled hate speech, which has predominantly linked online discourse solely to anti-Asian and anti-Chinese sentiments [9, 12]. Here, we find that racism itself can shift in meaning in line with wider changes in the global political context.

Our analysis further suggests several insights for the more general study of online hate speech. While state-of-the-art efforts at detecting hate speech have certainly been valuable [4, 6], our work demonstrates novel yet flexible and straightforward techniques for characterizing hate speech within its interactive contexts. More specifically, we show how hate speech predictions can signal different behaviors when it is prevalent among tweets which receive replies (in which case we identify interactions which call out racism) and when hate is present in replies themselves (in which case we detect the hateful use of the racist label itself). These characterizations are deepened in relation to their associated identity categories [8], since the former retaliates against racially charged xenophobia and the latter performs political criticism.

The most immediate next step in this line of work would be to examine the prevalence of bots in these datasets. Information operations may be involved in the spread of hate speech and in fueling the racism discussion at different points in time, with different tactical objectives [14]. Analysis of bot behavior may therefore deepen the findings presented here. It may also be interesting to explore other potential targets that may have been called out by the tweets mentioning racism. For example, it would be fruitful to look at the users who were mentioned in the tweets discussing racism. Sets of tweets from these mentioned users could be collected and run through the hate speech model to see if those users were indeed being more hateful than the average Twitter user.

References

1. Carley, K.M., Cervone, G., Agarwal, N., Liu, H.: Social cyber-security. In: International Conference on Social Computing, Behavioral-Cultural Modeling and

Prediction and Behavior Representation in Modeling and Simulation. pp. 389–394. Springer (2018)

2. Carley, L.R., Reminga, J., Carley, K.M.: Ora & netmapper. In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer (2018)
3. Chiriboga, D., Garay, J., Buss, P., Madrigal, R.S., Rispel, L.C.: Health inequity during the COVID-19 pandemic: a cry for ethical global leadership. The Lancet **395**(10238), 1690–1691 (2020)
4. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media (2017)
5. Devakumar, D., Shannon, G., Bhopal, S.S., Abubakar, I.: Racism and discrimination in COVID-19 responses. The Lancet **395**(10231), 1194 (2020)
6. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) **51**(4), 1–30 (2018), publisher: ACM New York, NY, USA
7. Huang, B., Carley, K.M.: Disinformation and misinformation on twitter during the novel coronavirus outbreak. arXiv preprint arXiv:2006.04278 (2020)
8. Joseph, K., Wei, W., Benigni, M., Carley, K.M.: A social-event based approach to sentiment analysis of identities and behaviors in text. The Journal of Mathematical Sociology **40**(3), 137–166 (2016)
9. Li, Y., Galea, S.: Racism and the COVID-19 epidemic: Recommendations for health care workers. American Journal of Public Health **110**(7), 956–957 (2020)
10. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology **54**(1), 547–577 (2003)
11. Reicher, S., Stott, C.: On order and disorder during the covid-19 pandemic. British Journal of Social Psychology **59**(3), 694–702 (2020)
12. Stechemesser, A., Wenz, L., Levermann, A.: Corona crisis fuels racially profiled hate in social media networks. EClinicalMedicine **23** (Jun 2020). https://doi.org/10.1016/j.eclinm.2020.100372
13. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology **29**(1), 24–54 (2010)
14. Uyheng, J., Carley, K.M.: Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines. Journal of Computational Social Science pp. 1–24 (2020)
15. Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., Carley, K.M.: Interoperable pipelines for social cyber-security: Assessing Twitter information operations during NATO Trident Juncture 2018. Computational and Mathematical Organization Theory pp. 1–19 (2019)
16. 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., et al.: Using social and behavioural science to support COVID-19 pandemic response. Nature Human Behaviour pp. 1–12 (2020)