Original Manuscript

Social Psychological and Personality Science
1–11

DOI: 10.1177/19485506211059329
journals.sagepub.com/home/spp
SAGE

Morally Homogeneous Networks and Radicalism

Mohammad Atari[1], Aida Mostafazadeh Davani[1], Drew Kogon[1], Brendan Kennedy[1], Nripsuta Ani Saxena[1], Ian Anderson[1], and Morteza Dehghani[1]

Abstract

Online radicalization is among the most vexing challenges the world faces today. Here, we demonstrate that homogeneity in moral concerns results in increased levels of radical intentions. In Study 1, we find that in Gab—a right-wing extremist network—the degree of moral convergence within a cluster predicts the number of hate-speech messages members post. In Study 2, we replicate this observation in another extremist network, Incels. In Studies 3 to 5 (N = 1,431), we demonstrate that experimentally leading people to believe that others in their hypothetical or real group share their moral views increases their radical intentions as well as willingness to fight and die for the group. Our findings highlight the role of moral convergence in radicalization, emphasizing the need for diversity of moral worldviews within social networks.

Keywords

morality, moral convergence, hateful rhetoric, radicalism, social networks

On January 6, 2021, a radicalized mob of rioters, spurred on by former President Trump's false allegations of a "rigged election," stormed the U.S. Capitol Building during a joint session of Congress, taking over the Senate Chamber and staff offices in an unprecedented insurrection which resulted in the death of five people (Safdar et al., 2021). Bolstered by former President Trump, the rioters openly organized on far-right social media websites—namely Gab and Parler—and arrived at the Capitol in response to calls for violence against Congress (Frenkel, 2021). What bonded the rioters was a shared belief that the election was stolen, and that it was their sacred duty to "stop the steal" by any means necessary, including intimidation and violence (Kunst et al., 2019).

Here, we investigate how such homogeneous moralized views in social networks may result in radicalism, and what role familial-like bonds with the ingroup play in this process. Specifically, we test the relationship between moral convergence and incendiary, hateful rhetoric in two extremist online social networks (Study 1: Gab; Study 2: Incels), and complement the findings with three controlled social-psychological experiments in a cross section of U.S. adults (Studies 3, 4, and 5) to further establish and dissect the link between moral convergence and radicalism.

Morally Motivated Violence

A pressing question across the social sciences has been to identify what factors lead normal people to become radicalized and engage in activities deemed as violating important social norms (Kruglanski et al., 2014). Recent theoretical advances have highlighted the roles of various social processes in radicalism, including placing significance and meaning on belonging to social categories, group-related narratives, and networks (Kruglanski et al., 2017).

Once an issue becomes moralized, people are more likely to act upon it to ensure their group prevails by nearly any means necessary, including persuasion, lying, cheating, protesting, or—in extreme cases—violence (Graham & Haidt, 2012; Skitka et al., 2021). Moral emotions such as anger, disgust, and hatred toward the outgroup are typically influential in "us vs. them" contexts, creating vehement tribal bonds with one's ingroup and extreme hostility toward outgroup members who may be dehumanized and seen as deserving harm and persecution. Fiske and Rai (2014) demonstrated this sense of moral righteousness across a wide range of contexts, characterizing it as "virtuous violence." Recent empirical work has started examining how radicalized individuals license themselves the moral

[1]University of Southern California, Los Angeles, USA

Corresponding Author:
Mohammad Atari, Department of Psychology, University of Southern California, 362 S. McClintock Ave, Los Angeles, CA 90089-161, USA.
Email: atari@usc.edu

authority to target other groups using extreme behaviors (Atran et al., 2007; Gómez et al., 2017).

More recently, it has also been demonstrated that the perception that one's moral views are shared by others—a phenomenon referred to as *perceived moral convergence*—increases the effect of moralization on the acceptability of violence and radical behavior (Mooijman et al., 2018). Specifically, in inter-group conflicts, the elicitation of violence appears to be often contingent on the moralization of the dispute (Finkel et al., 2020), and on the belief that other ingroup members share those moralized attitudes. Prior to the widespread adaptation of online social networks, perceived moral convergence would take place among a "band of brothers" or "soccer pals" (Atran, 2010, p. XII) where shared moralized principles would tighten the group's bonds, and the perception of the violation of those principles would result in radicalized intentions—even suicide bombings—against the outgroup.

Today, the perception of moral convergence is a frequently experienced phenomenon among social network users. Indeed, moral values—especially that of moral purity (e.g., religiosity, sacredness)—have been identified as a source of homophily and an important factor in the formation of bonds and communities online (Dehghani et al., 2016), often resulting in moral echo chambers with highly exaggerated perceptions of homogeneity among the members (Ashokkumar et al., 2020; Price et al., 2006).

The Present Research

In this research, we hypothesize that (a) users who are in a morally homogeneous information ecosystem are more likely to post hate speech to derogate and dehumanize the outgroup, and that (b) perceived moral convergence would entice pro-group radicalism intentions through identity fusion (see Swann et al., 2014) and a perceived sense of power (see Mooijman et al., 2018). We test these predictions in five studies with a diverse set of methodologies and operationalizations. In Study 1, we test the hypothesis that users in highly morally homogeneous clusters propagate more hate speech on Gab.com. Study 2 replicates this observation in a subreddit referred to as *Incels*. Study 3 examines whether leading people to believe they are in a morally convergent situation increases radicalism in laboratory settings. Study 4 replicates our findings in a nationally stratified sample in the United States and tests two possible concomitant mediating mechanisms in the causal link between perceived moral convergence and radicalism: identity fusion and sense of power. Finally, Study 5 demonstrates that the effect of moral convergence on radicalism can be observed in real, rather than made-up, groups, goes beyond mere alignment of nonmoral preferences with the group, and generalizes to other extreme pro-group inclinations.

Study 1

In Study 1, our goal is to investigate whether moral convergence in individuals embedded within social networks is associated with radicalization, captured by the frequency of hate-speech posts. We focus on Gab, as it claims to celebrate free speech and has attracted a large number of users who identify with far-right ideologies. Indeed, Gab played a major role as a hub for organizing the attempted insurrection on January 6, 2021 (Frenkel, 2021). We calculate the similarity of each user's moral concerns based on the typology of Moral Foundations Theory (MFT; i.e., care, fairness, loyalty, authority, and purity; see Graham et al., 2013), as measured through their language, to their cluster's average moral concerns to operationalize moral convergence. We consider this similarity score as the indicator for convergence and use it in a statistical model to predict the number of hate-speech posts by each user.

Method

To predict the moral and hateful content of posts, we first annotate a subset of the data according to expert-defined taxonomies of hate speech and moral language, train state-of-the-art classifier models on the annotations, and then predict labels for the remaining, unannotated data. We then conduct a cluster analysis on the network based on "follower" relations among Gab users. Finally, we conduct a multilevel model to predict the number of hate-speech posts as a function of users' moral convergence in their respective cluster.

Annotations. First, 7,692 messages posted by 800 randomly selected Gab users were manually annotated for hateful rhetoric and moral values based on previously developed coding manuals (Hoover et al., 2020; Kennedy et al., 2020). Specifically, seven expert annotators were trained to code "hate-based rhetoric" based on a coding guide developed according to a fusion of legal, computational, and social scientific perspectives on hate speech (see Kennedy et al., 2020). These elements guided annotators to identify rhetoric, beyond typical legal requirements of hate speech, that dehumanized, attacked, or advanced hatred toward a particular social group defined by race, religion, sexual orientation, nationality, gender, political affiliation, or mental or physical health status (Kennedy et al., 2020) (see Table 1). Similarly, the same data set was annotated for 10 moral labels, reflecting the vice and virtue dimensions of each of the 5 moral foundations in MFT (see Hoover et al., 2020). As specified in the coding manual, annotators identified linguistic queues (e.g., calling an event "unfair," saying that "they had no right to do that") that aligned with the corresponding category of MFT. In doing so, annotated labels help to discriminate moral uses of moral words from nonmoral uses (e.g., "turn right or left"). Inter-annotator

Table 1. Examples of Each Moral and Hate Labels in Gab Posts.

Label	Example
	Moral labels
Care	The only way to change things is to have compassion and being as rational as possible. The left vs right stuff isn't all life is about lol.
Harm	Check out this article proving that vaccines kill kids. [LINK]
Fairness	Seems only fair that lefties who enjoy harassing ICE agents at home get equal treatment.
Cheating	I worry my early vote will be somehow stolen, plus waiting until election day keeps the pollsters guessing and leaves little time to mount a counter offensive
Loyalty	Liberals are absolutely freaking out about this Boy Scout thing . . . I guess they're not used to seeing pride in your country and and pride in your President.
Betrayal	look imma be blunt here: FUCK "thoughts and prayers" for @SenJohnMcCain. Fuck McCain. He can rot in Hell, which is where he is absofuckinlutely going. Fucking traitor.
Authority	There's a New Sheri In Town and his name is; PRESIDENT DONALD J. TRUMP ! He was on POINT and on FIRE ! GOD BLESS PRESIDENT TRUMP ! #StopTheBias#MAGA#GabVets#GabFam
Subversion	When will the citizens of America wake up to the clear fact, they don't need Washington D.C. anymore.
Purity	This veteran could use some prayers right now; his life teeters in the balance as I type this out #PrayerRequest #PrayerWarriors #Powerof- Prayer #Prayer
Degradation	. . . When the truth is out that the Catholic church is nothing but a sex cult it will collapse and be no more.
	Hate-based rhetoric
Hate-based rhetoric	"'Texas' Couple Kept 5 Year Old African Girl as a Slave for 16 Years."'Texas couple' indeed. Meet Mohammed & Denise, just your typical Texans. How many times do these cases not turn out to involve somebody named Mohammed?

agreement coefficients for multiple annotators (Fleiss, 1971), after correction for the low prevalence of labels, prevalence- and bias-adjusted kappa (PABAK; Byrt et al., 1993) was computed for each of the 10 moral values categories and for the hate-based rhetoric label (PABAKs in Table 3). Examples from the current corpus are shown in Table 1.

Neural Network Text Classifiers. We trained one Long Short-Term Memory (LSTM) neural network model (Hochreiter & Schmidhuber, 1997) to predict whether posts contain hate speech. A second model, a multi-task LSTM network (Collobert & Weston, 2008), was trained to predict the presence of *binding* (i.e., loyalty, authority, and purity) and *individualizing* (i.e., care and fairness) moral foundations simultaneously. LSTMs incorporate a recurrent structure that encodes long-term dependencies between words and their past context. In both the single-task and multi-task models, posts were represented as sequences of pretrained GloVe word embeddings (Pennington et al., 2014) corresponding to the words in the original post. This embedding layer was then input to a 100-dimensional LSTM layer which is connected to a layer of fully connected units, with 0.33 dropout ratio (Srivastava et al., 2014). A sigmoid transformation is then applied to the output of the final layer to generate probabilistic predictions for the outcome. Models were trained to maximize the F_1 score on a validation set; Table 2 presents the precision, recall, and F_1 accuracy scores. After training and validation, the first model was used to automatically detect the presence of hateful rhetoric, and the second model to detect each moral vice/virtue in a large Gab corpus, consisting of 24,978,951 posts from 236,823 users, after removing non-English posts.

Community Detection Algorithm. Cluster analysis on the Gab social network was performed using the *Infomap* algorithm (Rosvall et al., 2009); a community detection algorithm for directed graphs with time complexity linear to the number of edges. This algorithm tries to detect communities by minimizing the description length for a random walk on the graph. Essentially, the probability flow of random walks within a network is being used as a proxy for the flow of information within the system, which is then used to break down the network. From the original Gab network data, users were filtered if they were not in our data of 24 million automatically labeled posts. This resulted in a user network consisting of 214,484 users, with approximately 16.9 million "follower" relations, consisting of one Gab user (the follower) that subscribes to the posts of another user (the "followee"). Infrequent users of Gab (less than 2 posts) and likely bots (at least 50,000 posts) were then pruned, leaving 15.6 million follower connections among 157,309 unique users. Infomap was applied to this Gab network graph. The resulting set of clusters ($n = 4{,}580$) were predominantly small ($Mdn = 3.0$ users), thus a threshold of 50 users was set for inclusion of a given cluster in the analysis. After applying this threshold, 126 clusters remained, accounting for 117,829 users ($Mdn = 74$).

Statistical Analysis. We capture the moral profile of the users by constructing a 10-dimensional vector, each cell

Table 2. First Row: Classification Metrics for Study 1, Trained and Tested on the Gab Data Set; Second Row: Classification Metrics for Study 2, Trained on the Gab Data Set, and Evaluated on the Incels Data Set (Study 2).

Corpus	Label	Precision	Recall	F_1 score
Study 1 (Gab)	Individualizing	0.74	0.63	0.68
	Binding	0.84	0.70	0.76
	Hate	0.51	0.72	0.60
Study 2 (Incels)	Individualizing	0.85	0.52	0.65
	Binding	0.87	0.51	0.64
	Hate	0.75	0.56	0.64

Note. Due to the sparsity of the 10 moral foundations in our data set, we investigate moral sentiment at a higher level (i.e., Individualizing and Binding). We aggregated care, harm, fairness, and cheating as the individualizing label, and loyalty, betrayal, authority, subversion, purity, and degradation as the binding label (see Table S1).

representing vice/virtue of each of the moral foundations, averaged from users' message history. The distance between a user and its cluster's moral profiles (averaged from all the posts of users in the cluster) was calculated as the Euclidean distance between the user profile and the user's cluster, which was then reversed to arrive at a convergence score. Subsequently, using the predicted labels for individualizing and binding moral values, we calculated each user's convergence based on their cluster's average scores, resulting in a binding and an individualizing convergence scores. A zero distance from a cluster's average was utilized to operationalize full moral convergence. These variables were then used as independent variables in a regression with users' count of hate-speech posts as the dependent variable. To account for the clustering information of the network, and the fact that many users had zero hate-speech posts, we ran a zero-inflated negative binomial mixed-effects model in which intercepts of distances were allowed to vary across clusters. We also offset our hierarchical models by the log transformation of total number of posts per person. Overall, 61,833 users (43.9%) had at least one hate-speech post.

Results

We first fit a negative binomial mixed-effects model with moral convergence predicting the number of their hate-speech posts offset by the total number of their posts. One standard deviation increase in moral convergence (based on the Euclidean distance) increased the rate of the hate-speech posts 30%, $b = 0.26$, $SE = 0.003$, $Z = 88.47$, $p < .001$. Therefore, it can be concluded that users who are closer to the moral average in a multidimensional space and share the majority's moral concerns in their cluster are more likely to post more hateful rhetoric on Gab.

In a follow-up analysis, we ran a similar model wherein users' convergence in binding and individualizing concerns, rather than general moral convergence, predicted the number of hate-speech posts. This model suggested that 1 standard deviation increase in the convergence of binding values increased the rate of the hate-speech posts by 33%, $b = 0.28$, $SE = 0.013$, $Z = 22.31$, $p < .001$. However, moral convergence in the individualizing values was not related to the rate of the hate-speech posts, $b = -0.02$, $SE = 0.013$, $Z = -1.28$, $p = .202$ (for robustness checks, see Supplementary Materials). Overall, these findings support the idea that, among Gab users, higher alignment of a user's moral language, especially binding moral language, with their cluster is associated with higher probability of using outgroup-derogatory, hateful language.

Study 2

In this study, we replicate the role of moral convergence in radicalized, hateful rhetoric in a different social network; a misogynist subreddit called "Incels" (founded for "involuntary celibates"). Most of the heated topics in this subreddit involve sexual topics and many members adhere to the "black pill" ideology, which espoused despondency often coupled with misogynistic views that condoned, downplayed, or advocated rape, while referring to women using dehumanizing language such as "femoids" and "sluts" (Jaki et al., 2019).

Method

We used a corpus of 906,455 comments by 34,165 unique users collected by Pushshift.io. First, to evaluate the accuracy of the classifier models used in Study 1 in assessing the language in the new corpus, four trained research assistants manually annotated 1,000 randomly selected comments for hate-based rhetoric (Kennedy et al., 2020) and moral values (Hoover et al., 2020). Inter-annotator agreement coefficients were again computed based on Byrt et al. (1993; PABAKs in Table 3). After assessing accuracy of the models developed in Study 1 on this annotated subset (Table 2), we used the models to detect moral values and hate speech on the whole data set.

Next, we removed users with less than five posts resulting in a data set with 11,454 users. Then, the trained models were applied to the remaining data set to determine hate and moral labels. As in Study 1, each user's moral profile

Table 3. Prevalence- and Bias-Adjusted Kappas (PABAKs) in Studies 1 and 2.

Dimension	Study 1	Study 2
Care	0.91	0.86
Harm	0.78	0.74
Fairness	0.88	0.61
Cheating	0.79	0.79
Loyalty	0.75	0.93
Betrayal	0.70	0.84
Authority	0.83	0.80
Subversion	0.83	0.55
Purity	0.93	0.68
Degradation	0.83	0.61
Hate	0.80	0.50

Note. Some of the PABAKs were relatively low for Study 2, hence Study 2's model may be less accurate than that of Study 1.

was calculated by averaging the moral values of all their posts. The average of all posts in this entire subreddit was also calculated as the moral profile of the community.[1] As in Study 1, a zero distance from the community's average was used to operationalize full moral convergence. The number of hate-speech comments were also calculated for each user in the data set. Overall, 10,240 (89.8%) had at least one hate-speech post.

Results

We first conducted a Poisson model with robust standard errors whereby moral convergence predicted the count of their hate-speech comments. Results indicated that 1 standard deviation increase in the moral convergence increased the rate of the hate-speech posts 3.94 times, $b = 1.37$, $SE_{Robust} = 0.124$, $Z = 391.60$, $p < .001$.

In a follow-up analysis, we ran a similar model, as in Study 1, whereby users' distance in individualizing and binding moral language predicted the count of hate-speech comments. This Poisson model suggested that 1 standard deviation increase in the convergence of binding values was associated with a 45% increase in the risk of posting hate speech, $b = 0.79$, $SE_{Robust} = 0.074$, $Z = 194.50$, $p < .001$. Users' convergence in the individualizing values was also positively associated with the count of hate-speech comments ($b = 0.70$, $SE_{Robust} = 0.126$, $Z = 173.5$, $p < .001$; for robustness checks, see Supplementary Materials). Results of this study suggest that the association between moral convergence and hate speech is not specific to Gab; rather, it replicates across platforms and is robust to different types of operationalizations and analyses. The higher the convergence between a user and their community in terms of expressed moral concerns, the more likely it is for them to use radicalized, hateful language.

Study 3

In Study 3, we examine the effect of perceived moral convergence on radicalism intentions in a controlled social-psychological experiment. Here, we experimentally manipulate participants' convergence (vs. nonconvergence) of their moral concerns with those of a hypothetical group on social media and assess subsequent radicalism intentions to protect this group.

Method

Participants and Procedure. We aimed to recruit 400 participants from Amazon's Mechanical Turk. After excluding participants who failed the attention check and those who left the main measure blank, the sample included 333 respondents. After agreeing to participate in the study, all participants were asked to choose a domain of moral concerns that was most important to them. There were five options in accordance with the five moral foundations (Graham et al., 2013). After they chose a moral concern, participants were randomly assigned to one of two conditions: *convergence* (vs. *nonconvergence*) in the selected moral concern with members of a Facebook group (see Supplementary Materials for more details). All participants completed the Inclusion of Self in Other Scale (ISOS; Aron et al., 1992) as a manipulation check.

Activism and Radicalism Intentions. The Activism and Radicalism Intentions Scale (ARIS; Moskalenko & McCauley, 2009) has two 4-item subscales for activism and radicalism. Each item was rated on a 5-point Likert-type scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*). We slightly reworded some items to be more consistent with our framework and manipulation. For example, the original item "I would join/belong to an organization that fights for my group's political and legal rights" was changed to "I would join/belong to an organization that fights for my group's moral beliefs." All coefficients are presented in Table 4.

Facebook Use. As we used a Facebook group as a part of our manipulation, we measured Facebook use to make sure that our results are not driven by participants' familiarity with the Facebook environment. All participants answered a single-item measure of Facebook use on a horizontal slider ranging from 0 (*Never*) to 100 (*All the time*).

Results

First, the manipulation checks were evaluated. Scores on the ISOS were substantially higher in the convergence condition ($t = 16.92$, $p < .001$, $d = 2.75$), suggesting that the experimental manipulation was successful. We ran a

Table 4. Internal Consistency Coefficients of Measures in Studies 3, 4, and 5.

Cronbach's alpha [95% CI]					
Study	Radicalism	Activism	Identity fusion	Power	Fight/die
Study 3	.73 [.68, .77]	.87 [.84, .89]	—	—	—
Study 4	.83 [.80, .85]	.88 [.87, .90]	.93 [.92, .94]	.85 [.82, .87]	—
Study 5	.72 [.69, .76]	.86 [.85, .88]	.93 [.92, .93]	.88 [.86, .89]	.86 [.84, .87]

Note. CI = confidence interval.

Welch-corrected t test to compare radicalism between the convergence and nonconvergence conditions. Participants who were assigned to the convergence group scored significantly higher on radicalism ($t = 2.90$, $p = .004$, $d = 0.32$; Figure 1). Results of the regression model indicated that after accounting for Facebook use and prioritized moral concern, moral convergence remained a significant predictor of radicalism ($B = 0.25$, $t = 2.67$, $p = .008$). We also found similar results for pro-group activism (see Supplementary Materials).

Study 4

In Study 4, we replicate and extend the effect of perceived moral convergence on radicalism intentions in a stratified nationally representative sample and examine two theoretically justified mediating mechanisms between them: identity fusion and sense of power.

Method

Participants and Procedure. After conducting a power analysis to detect a small effect ($f^2 = 0.02$) in a multiple regression model with a two-tailed p value of 0.05 and power of 90%, we aimed to recruit 500 participants in a national U.S. sample stratified across participants' gender, age, and political ideology. Participants were recruited through Qualtrics Panels. Our final sample included 510 American adults (254 male, 256 female) ranging in age from 18 to 70 ($M = 46.1$, $SD = 15.9$) and about half Democrats (49.7%) and half Republicans (50.3%). Participants were randomly assigned to convergence (vs. nonconvergence) condition, then completed a set of self-report measures, as in Study 3.

Our experimental manipulation of moral convergence was identical to that of Study 3.

Measures. We used the ARIS (Moskalenko & McCauley, 2009) as in Study 3. To assess the strength of a visceral feeling of oneness with the group participants were assigned to, we used the 7-item Identity Fusion Scale (Gómez et al., 2011). Finally, we used the 6-item version of the Sense of Power Scale (Anderson et al., 2012). All coefficients were high (see Table 4).

Results

Comparing the manipulation check between the conditions suggested that participants in the convergence condition scored substantially higher than the nonconvergence condition on inclusion of self in others ($t = 10.78$, $p < .001$, $d = 0.95$), suggesting that the experimental manipulation was successful. We first conducted a Welch-corrected t test to investigate the main effect of perceived moral convergence on radicalism. Mirroring Study 3's main effect, we found radicalism intentions to be significantly higher in the convergence (vs. nonconvergence) condition, $t = 2.66$, $p = .008$, $d = 0.24$. We then conducted a structural equation model (SEM), with moral convergence as the predictor and two mediators to predict radicalism as shown in Figure 2. Moral convergence did not have a direct effect on radicalism ($c' = 0.02$, $SE = 0.08$, $p = .832$) after accounting for mediators. Identity fusion mediated the relationship between moral convergence and radicalism ($ab = 0.18$, $SE = 0.04$, bootstrap $p < .001$); however, the mediating effect of sense of power was not significant ($ab = 0.04$, $SE = 0.03$, bootstrap $p = .240$). Adjusting for covariates did not change the total effect substantially (Supplementary Materials). We also found consistent results for pro-group activism (see Supplementary Materials).

Study 5

In this experiment, we address three limitations of Studies 3 and 4. Specifically, we aim to show that (a) the effect on radicalism is specific to "moral" convergence, not other types of amoral homogeneity in attitudes; (b) the effect can be observed in "real" groups rather than hypothetical groups; and (c) the effect is not limited to radicalism intentions and can be generalized to outcomes related to extreme pro-group acts, such as the willingness to fight and die for the ingroup. Based on previous studies, we hypothesized that moral convergence with other Americans increases pro-U.S. radicalism and the willingness to fight and die for American values. Accordingly, we expected an interaction

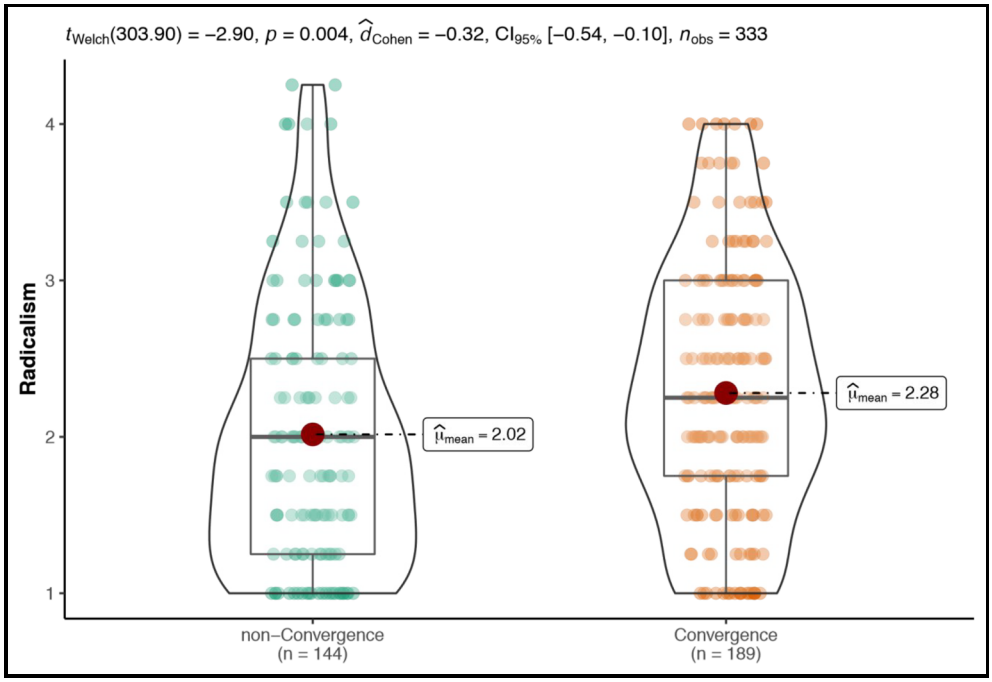

Figure 1. Radicalism Intentions In Morally Convergent and Nonconvergent Conditions (Study 3).

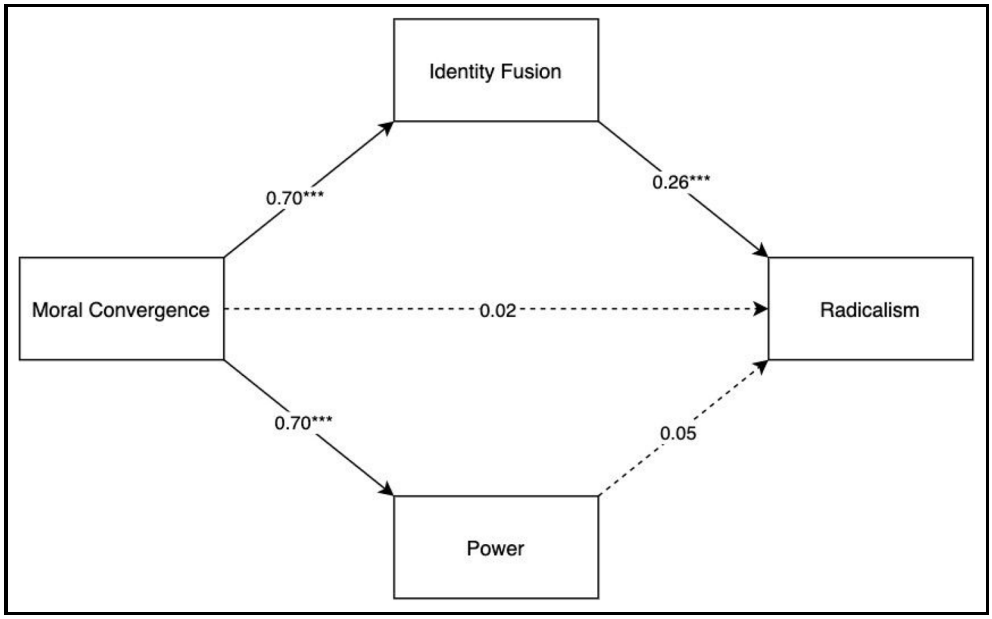

Figure 2. The Path Model of the Relationship Between Moral Convergence And Radicalism (Study 4). Covariance Between Identity Fusion and Power is Not Shown.
*$p < .05$. **$p < .01$. ***$p < .001$.

between domain (moral vs. nonmoral) and convergence (vs. nonconvergence) in predicting these outcomes.

Method

Based on the same specification of power analysis in Study 4 ($f^2 = 0.02$, $p = .05$, power = 90%), we recruited 588 adults born and currently living in the United States (239 male, 339 female, 10 nonbinary), ranging in age from 18 to 77 ($M = 37.8$, $SD = 13.2$). Participants were randomly assigned to each condition in a 2 (convergence vs. nonconvergence) × 2 (moral vs. nonmoral) between-subjects design (see Supplementary Materials). First, participants were asked about their moral (vs. nonmoral) attitudes.

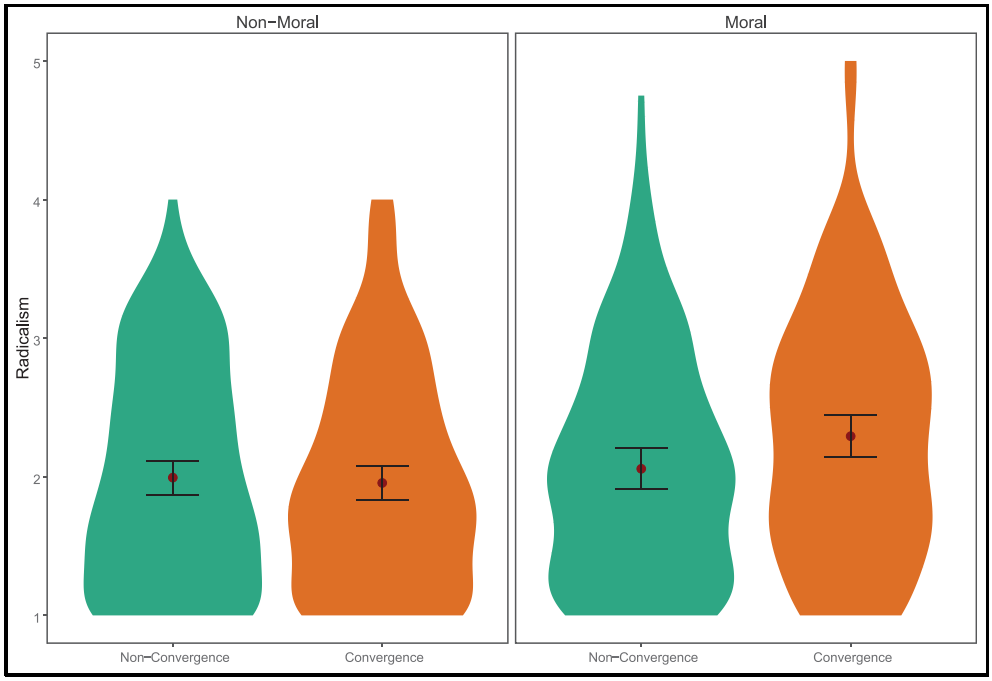

Figure 3. Radicalism Intention Scores as a Function of the Type of Manipulation (Study 5).

Then, in the convergence (vs. nonconvergence) condition, participants were told that "According to our national database, most (vs. a small number of) Americans share your interest in . . ." Then participants completed the state version of Identity Fusion Scale (Gómez et al., 2019), the state version of the Sense of Power Scale (Anderson et al., 2012), ARIS (Moskalenko & McCauley, 2009), and an adapted version of the 7-item measure of willingness to fight and die (Swann et al., 2009; internal consistency coefficients are shown in Table 4).

Results

Consistent with our hypotheses, in a two-way analysis of variance (ANOVA), there was a significant interaction between morality and convergence in predicting people's radicalism intentions to protect American values (F = 4.01, p = .046, η_p^2 = 0.007). In addition, the main effect of moral (vs. nonmoral) condition was significant (F = 8.78, p = .003, η_p^2 = 0.01). Post hoc tests indicated that participants' radicalism intentions were higher in the moral convergence condition (M = 2.29, SD = 0.89) than in the nonmoral nonconvergence (M = 1.99, SD = 0.79, difference = 0.30, 95% confidence interval (CI) = [0.05, 0.55], p_{adj} = .010), nonmoral convergence (M = 1.96, SD = 0.79, difference = 0.34, 95% CI = [0.09, 0.58], p_{adj} = .003), and moral nonconvergence (M = 2.06, SD = 0.83, difference = 0.24, 95% CI = [−0.03, 0.50], p_{adj} = .098) conditions (see Figure 3). The interaction of morality and convergence was robust to inclusion of political conservatism as a covariate (F = 4.36, p = .037, η_p^2 = 0.007). A follow-up linear model with the same independent variables suggested that political conservatism was negatively associated with radicalism intentions (B = −0.04, SE = 0.02, p = .027).

A two-way ANOVA to predict people's willingness to fight and die for the United States and its values yielded consistent but smaller effects (see Supplementary Materials). These results collectively show that the effect on radicalism is specific to moral convergence, not other types of amoral homogeneity or heterogeneity in attitudes. We also demonstrate that the effect can be observed in "real" groups (here, nationality) rather than hypothetical groups, although we acknowledge that manipulating moral convergence may be a difficult task in some real groups when one knows a great deal about others. Finally, the effect is not limited to radicalism intentions and can be generalized to outcomes related to extreme pro-group acts, such as the willingness to fight and die for the United States and the values it stands for.

General Discussion

Two days before his account would be permanently suspended "due to the risk of further incitement of violence,"[2] former President Trump tweeted "These are the things and events that happen when a *sacred* landslide election victory is so *unceremoniously* & *viciously stripped away* from great *patriots* who have been badly & *unfairly* treated for so long [emphases added]." Brady et al. (2020) have recently demonstrated the efficacy of the circulation of such

moralized rhetoric in online echo chambers. We argue, in this work, that this type of rhetoric is further validated and reinforced in the congenial atmosphere of social networks, creating a perception of moral homogeneity, and a moral obligation to defend the ingroup even by radical means, as it transpired in the U.S. Capitol on January 6, 2021.

In Study 1, we found that users who converged with their cluster's moral profile were more likely to disseminate hate speech, language intended to dehumanize, or call for violence, against outgroup members. In Study 2, we successfully replicated this effect in Incels, a social network in which members disseminated misogynistic rhetoric. These observational studies point to the possibility that homogeneity in moral worldviews within social networks could potentially result in validation and reinforcement of the common attitudes, and consequently, in radicalized behaviors. No direction of causality, though, could be inferred from these observational social-network studies (DellaPosta et al., 2015; Shalizi & Thomas, 2011).

Our follow-up studies help us better understand the mechanisms explaining the link between moral convergence and radicalism in controlled experimental settings. Study 3's results indicated that when people were led to believe that their group was morally homogeneous, they became more inclined to act in favor of the group, and even willing to commit radical acts to protect the group. In Study 4, we replicated the previous study in a stratified sample of American adults, and demonstrated that identity fusion with the ingroup (see Swann et al., 2009), but not perceived sense of power, explained the relationship between moral convergence and radicalism. Therefore, when individuals perceive homogeneity of moral views in their group, they develop a visceral sense of oneness with their group, which in turn can radicalize them into being willing to commit outgroup-derogatory acts to protect and preserve their own group (Atran, 2021; Gómez et al., 2017). Study 5 provided experimental evidence that moral convergence (and not amoral convergence) has a unique effect on radicalism and willingness to fight and die for a real ingroup, that goes beyond mere moralization and amoral attitude homogeneity.

By highlighting the role of moral convergence and identity fusion, the current research has important practical implications. Our results highlight the importance of moral diversity in online social networks to avoid affective polarization and creation of moral echo chambers that could contribute to radicalization through formation of cult-like identities to which individuals get vehemently attached. Specifically, it is necessary for deradicalization efforts (Johnson et al., 2019) to diversify morally homogeneous information ecosystems by attitude "demoralization" (Skitka et al., 2021) and encouraging "defusion" from the group (Fredman et al., 2015).

Finally, a constraint on generality of these findings is worth mentioning for replication and follow-up studies (Simons et al., 2017). First, our experiments include only Western participants, hence our results may not be generalized to other cultures. Our computational studies are also mostly centered around U.S. politics, our language data are in English, and our model potentially embeds prediction biases that have recently been discovered in evaluating neural-network models. Second, using cross-sectional data for mediation analysis leads to biased estimates even under ideal conditions and, when used for examining the mechanisms underlying the effects of experimental manipulations, relies on the stringent assumption that the tested mediators are the only existing mediators (Fiedler et al., 2011). Hence, future studies are encouraged to test these models using longitudinal or experimental mediation designs.

Authors' Note

The data and materials used in this work are available at https://osf.io/q7f5r/

Acknowledgments

We are grateful to Merrick Osborne and Nils Karl Reimer for providing helpful feedback on an earlier version of the manuscript.

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 research was sponsored by NSF CAREER BCS-1846531 to MD.

ORCID iDs

Mohammad Atari https://orcid.org/0000-0002-4358-7783
Morteza Dehghani https://orcid.org/0000-0002-9478-4365

Supplemental Material

The supplemental material is available in the online version of the article

Notes

1. As subreddits do not give follower relations among users, we were unable to repeat the cluster analysis from Study 1. Thus, we treat the subreddit as a single community.
2. https://blog.twitter.com/en_us/topics/company/2020/suspension.html

References

Anderson, C., John, O. P., & Keltner, D. (2012). The personal sense of power. *Journal of Personality*, *80*, 313–344.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*, 596–612.

Ashokkumar, A., Talaifar, S., Fraser, W. T., Landabur, R., Buhrmester, M., Gómez, A., Paredes, B., & Swann, W. B. (2020). Censoring political opposition online: Who does it and why. *Journal of Experimental Social Psychology*, *91*, 104031.

Atran, S. (2010). *Talking to the enemy: Faith, brotherhood, and the (un) making of terrorists*. Harper Collins.

Atran, S. (2021). Psychology of transnational terrorism and extreme political conflict. *Annual Review of Psychology*, *72*, 471–501.

Atran, S., Axelrod, R., & Davis, R. (2007). Sacred barriers to conflict resolution. *Science*, *317*, 1039–1040.

Brady, W. J., Crockett, M., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*, 978–1010.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*, 423–429.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167). Association for Computing Machinery.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., Vaisey, S., Iliev, R., & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, *145*, 366–375.

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, *120*, 1473–1511.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*, 1231–1236.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., Mcgrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*, 533–536.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.

Fredman, L. A., Buhrmester, M. D., Gómez, A., Fraser, W. T., Talaifar, S., Brannon, S. M., & Swann, W. B. (2015). Identity fusion, extreme pro-group behavior, and the path to defusion. *Social & Personality Psychology Compass*, *9*, 468–480.

Frenkel, S. (2021, January 6). The storming of Capitol Hill was organized on social media. *The New York Times*. https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html

Gómez, A., Brooks, M. L., Buhrmester, M. D., Vázquez, A., Jetten, J., & Swann, W. B. (2011). On the nature of identity fusion: Insights into the construct and a new measure. *Journal of Personality and Social Psychology*, *100*, 918–933.

Gómez, A., López-Rodríguez, L., Sheikh, H., Ginges, J., Wilson, L., Waziri, H., Vázquez, A., Davis, R., & Atran, S. (2017). The devoted actor's will to fight and the spiritual dimension of human conflict. *Nature Human Behaviour*, *1*, 673–679.

Gómez, A., Vázquez, A., López-Rodríguez, L., Talaifar, S., Martínez, M., Buhrmester, M. D., & Swann Jr, W. B. (2019). Why people abandon groups: Degrading relational vs collective ties uniquely impacts identity fusion and identification. *Journal of Experimental Social Psychology*, *85*, 103853.

Graham, J., & Haidt, J. (2012). Sacred values and evil adversaries: A moral foundations approach. In M. Mikulincer & P. R. Shaver (Eds.), (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 11–31). American Psychological Association.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., & Dehghani, M. (2020). Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*, 1057–1071.

Jaki, S., De Smedt, T., Gwóüdü, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, *7*, 240–268.

Johnson, N., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., & Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*, 261–265.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Olmos, G., Omary, A., Park, C., Wijaya, C., Wang, X., . . . Dehghani, M. (2020). The gab hate corpus: A collection of 27k posts annotated for hate speech. *Psyarxiv*. https://doi.org/10.31234/osf.io/hqjxn

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, *35*, 69–93.

Kruglanski, A. W., Jasko, K., Chernikova, M., Dugas, M., & Webber, D. (2017). To the fringe and back: Violent extremism and the psychology of deviance. *American Psychologist*, *72*, 217.

Kunst, J. R., Dovidio, J. F., & Thomsen, L. (2019). Fusion with political leaders predicts willingness to persecute immigrants and political opponents. *Nature Human Behaviour*, *3*, 1180–1189.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*, 389–396.

Moskalenko, S., & McCauley, C. (2009). Measuring political mobilization: The distinction between activism and radicalism. *Terrorism and Political Violence*, *21*, 239–260.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the empirical methods in natural language processing (emnlp 2014)* (*Vol. 12*, pp. 1532–1543). Association for Computational Linguistics.

Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. *Communication Theory*, *16*, 47–74.

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, *178*, 13–23.

Safdar, K., Ailworth, E., & Seetharaman, D. (2021, January 8). Police identify five dead after capitol riot. *The Wall Street Journal*. https://www.wsj.com/articles/police-identify-those-killed-in-capitol-riot-11610133560

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*, 211–239.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. *Annual Review of Psychology*, *72*, 347–366.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*, 1929–1958.

Swann, W. B., Buhrmester, M. D., Gómez, A., Jetten, J., Bastian, B., Vázquez, A., Ariyanto, A., Besta, T., Christ, O., Cui, L., Finchilescu, G., González, R., Goto, N., Hornsey, M., Sharma, S., Susianto, H., & Zhang, A. (2014). What makes a group worth dying for? identity fusion fosters perception of familial ties, promoting self-sacrifice. *Journal of Personality and Social Psychology*, *106*, 912–926.

Swann, W. B., Gómez, A., Seyle, D. C., Morales, J., & Huici, C. (2009). Identity fusion: The interplay of personal and social identities in extreme group behavior. *Journal of Personality and Social Psychology*, *96*, 995–1011.

Author Biographies

Dr. Mohammad Atari completed his Ph.D. in the Department of Psychology at the University of Southern California, and is currently a Postdoctoral Fellow at Harvard University. His broad research interest is why and how morality binds people together, but also blinds them into "us" vs. "them." His current research examines moral values and cultural change taking a cultural evolutionary perspective, using a collection of methodological approaches including social psychological experimentation and Natural Language Processing (NLP).

Aida Mostafazadeh Davani is a PhD candidate in Computer Science at the University of Southern California. Her primary research interests are Natural Language Processing (NLP) and ethics of Artificial Intelligence. Her current work examines the role of annotator bias (e.g., implicit and explicit stereotyping) in development of large, annotated datasets which are used in downstream NLP applications.

Drew Kogon is a graduate student in the Department of Psychology at the University of Southern California.

Brendan Kennedy is a 6th-year computer science PhD student at the University of Southern California, where he applies Natural Language Processing and Machine Learning to address questions in social psychology. His research focuses primarily on the relationship between morality and language as well as the annotation of psychological phenomena in text.

Nripsuta Ani Saxena is a PhD student in Computer Science at the University of Southern California. Her research interests are, broadly, fairness in machine learning, AI for Social Good, and questions at the intersection of psychology, sociology, and computer science.

Ian Anderson (he/they) is a third year Ph.D. student working under Professor Wendy Wood at the University of Southern California. His current research interests include how psychological and interpersonal processes are influenced by social media, technological design, and human-computer interactions. Specifically, his work examines habits, misinformation, conspiracies and rumors, identity and stereotypes, and how the designs of social media platforms influence user behavior through reward-learning processes.

Dr. Morteza Dehghani is an Associate Professor of Psychology and of Computer Science. He directs the Morality and Language Lab at the University of Southern California, where he bridges cutting-edge methods from Natural Language Processing (NLP) with social psychology. Most of his work focuses on morality and culture, as well as Fairness, Accountability and Transparency in machine learning.

Handling Editor: Igor Grossmann