PLOS Complex Systems

RESEARCH ARTICLE

Online identity and action discourse around the 2020 and 2024 U.S. presidential elections

Mikhail Lipatov[1*], **Lucia Illari**[2], **Richard Sear**[2], **Akshay Verma**[2], **Neil F. Johnson**[2], **Sergey Gavrilets**[1,3,4,5]

1 Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, United States of America, **2** Physics Department, George Washington University, Washington, District of Columbia, United States of America, **3** Department of Mathematics, University of Tennessee, Knoxville, Tennessee, United States of America, **4** Complexity Science Hub, Vienna, Austria, **5** Institute for Advanced Study in Toulouse, Toulouse School of Economics, Toulouse, France

* mlipatov@utk.edu

OPEN ACCESS

Citation: Lipatov M, Illari L, Sear R, Verma A, Johnson NF, Gavrilets S (2026) Online identity and action discourse around the 2020 and 2024 U.S. presidential elections. PLOS Complex Syst 3(5): e0000107. https://doi.org/10.1371/journal.pcsy.0000107

Editor: Luca Maria Aiello, IT University of Copenhagen: IT-Universitetet i Kobenhavn, DENMARK

Received: October 31, 2025

Accepted: April 2, 2026

Published: May 13, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcsy.0000107

Abstract

Social psychology theories of collective action argue that shared social identity mediates coordinated behavior through interacting collective mental states, such as the perception of a common grievance, the corresponding corrective action norms, and the associated efficacy beliefs. Large-scale social media data make it possible to test these theories quantitatively, but operationalizing them in online settings remains challenging. We propose an operationalization that maps the theoretical mental states onto discourse and evaluates their temporal interdependence using trend analysis, stationarity tests, vector autoregression, and pathway analysis. We apply this framework to the communal discussion of alleged voter fraud in the ~90 million social media posts around the 2020 and 2024 U.S. presidential elections. In 2020/2021, we observe a canonical interaction sequence that the theories suggest: grievance about the alleged electoral fraud predicts subsequent mutual validation, validation predicts shared identity, and identity predicts efficacy beliefs and action discourse. Overall, the results are consistent with a collective psychological alignment that strengthens in the runup to January 6, 2021. Conversely, the 2024/2025 results do not reliably support either the alignment or the canonical sequence. Instead, the relationships are often negative or weak: grievance can suppress efficacy, action can reduce grievance, and identity predicts validation without consistently predicting action. These contrasts show that the coupling among grievance, validation, identity, efficacy, and action in digital conversation is context-dependent rather than universal. By operationalizing psychological theories in online discourse, the study both confirms specific theoretical mechanisms behind collective action in one electoral context and identifies the conditions under which the mobilizing alignment fragments in another.

Data availability statement: The code that is necessary to replicate this article's analysis is available as a record on Zenodo (https://doi.org/10.5281/zenodo.18759831). The record also includes the numeric time series that we derive from filtering the posts -- the time series that enter into the auto-regression analyses. Apart from several demonstrative examples in the Supplementary Text, we do not publish the text of individual social media posts that we analyze. Nor do we publish any other information on the individual posts, such as their timestamps or communities.

Funding: M.L., L.I., N.F.J., S.G., R.S. and A.V. were supported by the John Templeton Foundation grant 62434; templeton.org. S.G. was funded by the US Air Force Office of Scientific Research (grants FA9550-21-1-0217, FA9550-22-1-0250, FA9550-20-1-0382 and FA9550-20-1-0383; afrl.af.mil/AFOSR/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

We adapt the existing theories of social identity formation and development to large-scale social media data. In connecting the theories with the data, we utilize standard time-series statistical methodologies. We apply our framework to the development of the social identity that arose from the discussion of electoral fraud in the wake of the 2020 U.S. presidential election on the right-wing extremist social media networks. This social identity potentially served as the foundation for the subsequent attack on the U.S. Capitol on January 6, 2021. Our findings support this hypothesis and the canonical developmental sequence wherein the mutual validation of a discussed common grievance leads to the formation of an associated common identity that, in turn, elicits calls for corrective action. On the other hand, this identity development does not occur around the 2024 election, probably due to the lack of the necessary stimulating societal factors during this later time.

Introduction

Collective action is a central driver of social and political change, from the struggles for civil rights to the recent global protest waves. In the twenty-first century, mobilization increasingly originates in digital spaces, where grievances are voiced and identities take shape in real time. Examples such as the Arab Spring uprisings [1], the Occupy movement [2], and transnational solidarity with refugees following the death of Aylan Kurdi [3,4] highlight the transformative power of online communication to enable collective consciousness and coordinated action.

Online activism often owes its effectiveness to specific mechanisms and strategies that amplify grievances and elicit action in the context of platform-mediated power relations. For example, climate activists perceive social-media platforms as commercially driven and adapt through "faking, optimising and conceding to power" [5]. Along the same general lines, participants in Hong Kong's Anti-ELAB movement utilize digital media to co-produce political consumerism via the integration of symbolic resources, responses to ongoing events, and the attraction of mainstream media and public figures' attention [6]. Increased coordination of online media messages using specific hashtags helps channel political sentiment into calls for offline protest and mobilization [7]. Several recent studies further highlight the strategic and mechanistic dimensions of mobilization. For example, relationship-building with regime pillars can be important for campaigns that resist democratic backsliding [8], while artivism balances internal solidarity with external outreach [9].

Emotions and identity formation online are also frequently critical for offline mobilization. In particular, non-beneficiary participation in the Black Lives Matter movement was driven by pandemic-induced compassion and guilt [10], while ethnography of the Yellow Vests points to a reactive identity forged through shared stigma and rejection

by authorities [11]. Thus, expressions of grievance are often linked to perceived victimhood, moral righteousness, feelings of dispossession and distrust of institutions.

A number of theoretical frameworks seek to model the social processes that lead to collective action via the above-mentioned strategic and emotional mechanisms. One set of frameworks derives the social dynamics from the distributions of personal preferences and subjective utilities of action [12–17]. Other theoretical developments focus on the way that the spread of behavior is affected by the patterns of informational connections in social networks [18–23].

The social psychology of identification, in particular, can provide a strong theoretical foundation for understanding the processes that lead to collective behavior. Social Identity Theory (SIT) [24] and Self-Categorization Theory (SCT) [25] explain how individuals derive self-definition and motivation from group memberships. Whereas these frameworks specify *why* group identities matter, they are less precise about *how* shared identities emerge from communication, as well as how online expressions of grievance, identity, and efficacy coalesce into action. The empirical work that addresses this gap has explored these developmental processes in case studies.

Such empirical work has shown that the shared identities that form the foundations for collective action are actively constructed and reinforced through communicative interaction. Intragroup interaction provides the arena for consensus formation and validation, whereby subjective beliefs become collectively endorsed standards. Festinger's [26] theory of social comparison highlights the role of communication in providing subjective certainty. Such certainty aids the active construction of perceived group realities and the formation of shared cognitions such as stereotypes [27]. Group interaction can generate new shared identities and norms even in the absence of pre-existing categories [28,29].

Historical cases illustrate these processes. During the Arab Spring, the circulation of images and reports via social and satellite media created consensus around the illegitimacy of regimes, enabling the emergence of opinion-based groups united by shared demands for change [1]. In the Occupy movement, participants in online forums converged both on grievances against inequality and on the normative action of "occupation," forming what Smith and coauthors [2] call an identity–norm nexus (INN). The viral circulation of Aylan Kurdi's image generated outrage and compassion, as online discussions transformed emotional responses into enduring solidarity with refugees [3,4]. In each case, communication produced consensus, validation, and normative clarity, enabling identity to become the basis for collective action.

While consensus and validation are necessary, they are not sufficient for collective action. Research highlights the importance of efficacy beliefs: the belief that collective effort can produce change. The Social Identity Model of Collective Action (SIMCA) [30] posits that injustice, identity, and efficacy jointly drive mobilization. Empirical studies confirm that efficacy beliefs are often forged and reinforced in communication, serving as tipping points that transform grievances into concrete action intentions [31,32].

Group interaction can also shape the trajectory of mobilization toward politicization or radicalization. Thomas and coauthors [33] demonstrated experimentally that discussion can produce either conventional, legal pathways of action (politicization) or more extreme, extra-legal orientations (radicalization), depending on whether radical strategies are validated as legitimate. This distinction resonates with field studies of both protest and extremism, where communication not only consolidates shared grievances but also defines the boundaries of acceptable action.

Recent digital-trace studies underscore these points. In a recent publication, Smith and coauthors [34] showed that online posts predicting offline protest attendance were distinguished less by ideological content than by validation (e.g., likes, retweets) and the presence of efficacy-related talk such as logistics and calls to mobilize. Wischerath and colleagues [35] documented how conspiracy-oriented Telegram networks circulated grievances and planning cues alongside violent messages, with network centrality magnifying their mobilizing force. Brown, Smith, and colleagues [36,37] compared extremists who mobilized to violence with those who did not, finding that mobilization was associated not with expressions of hate *per se*, but with communication about violent action, logistics, and planning. Together, these findings highlight the centrality of efficacy and validation in building a bridge between grievance expression and actual mobilization.

Identity formation is shaped by both top-down and bottom-up dynamics. Top-down processes include leadership cues, institutional structures, and iconic events. For example, the image of Aylan Kurdi provided a powerful top-down symbol that crystallized compassion into solidarity [3]. Similarly, elite calls for protest during the Arab Spring gave resonance to grassroots demands [1].

Bottom-up processes are equally crucial: communication among peers validates grievances, consensualizes norms, and generates new opinion-based groups. Occupy exemplified this dynamic: though sparked by Adbusters' call, the movement's identity and demands were co-created in online discussion [2]. Haslam and coauthors [38] likewise demonstrated how prisoners forged a resistant identity through interaction, even within constraining institutional structures.

Importantly, these dynamics reinforce one another. Elite cues gain traction only when validated within grassroots discussion, while grassroots talk is amplified when echoed by leaders and media. Social identity is thus best understood as the product of a recursive interaction between top-down framing and bottom-up validation.

Several of the above-mentioned recent studies have operationalized, tested and extended the classical psychological theories of identity formation such as SIMCA and SIT in observational social media data [1,2,34–36]. Such work complements the traditional accounts of social identity development and collective action by making use of a fundamentally new informational source. In other words, the data underlying the classical models is different from the online data in several important ways. For example, the SIMCA, EMSICA and normative alignment models originally define collective mental states in terms of data from laboratory experiments, surveys, scenario studies, and live field studies [30,31]. Similarly, the original formulations of the SIT and SCT models are primarily based on controlled laboratory experiments, which were supplemented by qualitative theoretical interpretations of real-world intergroup conflict and prejudice [25,39]. In contrast with the classical model data, social media data is both more abundant and less subject to experimental control. Accordingly, the best way to operationalize the psychological theories in online data is far from obvious, so that the benefit of applying classical theories to web-based information also constitutes the central challenge of this approach.

At the current stage of this research, the optimal operationalization may well depend on the data set under scrutiny. Nevertheless, as these research efforts continue, the scientific community will tend to a methodological consensus. The present study offers a new operationalization and a new data set that contribute to the emerging connection between classical theories of social identity formation and social media data. Several specific quantitative characteristics of our work make it valuable in this context. The data set is relatively extensive in terms of the number of posts, the time period, and the number of platforms. Furthermore, the operationalization is simultaneously well-suited for this data and incorporates the components of the above-mentioned theories in a simple, intuitive, modular, and flexible fashion. In the context of this operationalization, we find that our data set largely supports the existing psychological theories. Our analysis also suggests additions to these theories and their refinement.

Our focus is the sub-identity of the broader "MAGA" movement that coalesced around the alleged 2020 U.S. presidential election voter fraud. We test whether discourse dynamics are consistent with mechanisms that can precede mobilization; we do not claim that the January 6 attacks were caused by any specific individuals. To trace the identity's development, we code millions of posts for the relevant expressions of grievance, verbal markers of shared identity, expressions of proposed normative action, expressions of belief in this action's efficacy, and expressions of mutual validation. This allows us to track the co-evolution of these constructs in digital traces. Our analyses show how voiced grievances elicit validation, how efficacy expressions consolidate the shared identity, and how grievances, norms and identity markers align through communication and collective action. Our data reveal progressive normative alignment: validation and identity reinforcement produce increasingly convergent expressions of who "we" are, as well as what "we" can and should do together.

In this way, the study demonstrates how classic mechanisms identified in laboratory and case-study research – consensus, norm alignment, validation, efficacy beliefs, and identity formation – are observable in digital communication at

scale. By bridging social psychological theory with computational social science, we advance the understanding of how collective action emerges in contemporary communication environments.

We focus on the patterns of online discourse. We do not adjudicate the truth of any claims with regard to voter fraud, assign responsibility for offline events, or evaluate the morality or legality of actions. References to "MAGA" and related terms are operational text labels derived from expression markers. We do not associate any specific characteristics with the self-identified supporters of the "MAGA" movement. Inclusion in our dataset is content-based using a predefined hate-speech criterion, not political alignment.

Results

Unless otherwise stated, the following results pertain to the analysis of the 2020/2021 social media data. The analytical procedure for the 2024/2025 dataset is very similar; we describe the differences between the two procedures toward the end of this section.

Expression categories

We hypothesize that around the time of the 2020 U.S. presidential election, a new sub-identity emerged within the Make America Great Again (MAGA) movement. This sub-identity centered on beliefs in election fraud and frequently advocated extra-institutional action; these discussions temporally coincided with the period leading up to the January 6, an event not necessarily endorsed by the broader MAGA movement.

We aim to investigate the development of this sub-identity, its dynamics over time, and the conditions that contributed to its emergence and influence within the larger movement. To do so, we employ an empirical approach that resembles that of Smith and coauthors [2,40]. Specifically, we focus on the social media posts that are relevant to the formation of this identity, classifying them via computer regular expressions into several categories of expression.

1. *Grievance* talk – expression of the relevant grievance (i.e., the voting and election fraud);
2. *Action* talk – advocacy for action that is meant to correct the grievance (e.g., an impactful gathering of people in protest of the fraud);
3. *Efficacy* talk – expression of belief in the efficacy of such action (e.g., that such a gathering is likely to overturn the election's result);
4. *Identity* talk – expression of belonging to a collective identity that is associated with this grievance and this normative action (e.g., distinctive acronyms such as MAGA).
5. *Validation* talk – conversation that begins with expression that belongs to one of the above-mentioned four categories – *grievance*, *identity*, *action*, or *efficacy* – and is later followed by a response containing a validation-related regular expression (see the Supplementary Materials).

In the rest of the article, an italicized category name generally refers to the category itself or the number of posts in the category as identified by the regular expressions, versus a more abstract concept associated with the category. For example, *identity* can refer to the proportion of posts that we have obtained via this category's filters. The un-italicized word "identity", on the other hand, refers to the abstract concept of social identity that is the general topic of this article.

The regular expressions that we use to identify the above-mentioned categories of expression (see Materials and Methods, as well as the Supplementary Text in S1 File) are the first step in our operationalization of the existing social identity theories. In taking this step, we propose that the number of posts that these regular expressions select for each category reflects the collective mental state that corresponds to this category and that the theories seek to capture. We acknowledge that the correspondence between the theories and our operationalization may not be exact. For example,

our *identity* regular expressions track a social identity that existed before the articulation of the relevant grievance and was subsequently co-opted for a sub-identity to mobilize collective action in response to the grievance. This is in contrast with the idea of an entirely new social identity that appears in response to the grievance.

Time series

We analyze ~60 million posts in our "hate universe" database [41–43] (see also Materials and Methods), between Super Tuesday (March 3, 2020) and the Senate report on the Capitol attacks (June 8, 2021).

On each day between these two dates, we determine the fraction of posts that falls into one of the categories above. These time series are shown in Fig 1. We refer to the period before the election as Period 1, the period between the election and the attack as Period 2, and the period after the attack as Period 3. We scale each variable to have a mean of zero and standard deviation of one across all the time points in a given period (see Materials and Methods).

The existing social identity theories imply dynamic interactions among collective mental states. Our scaled time series are thus a natural second step in operationalizing these theories, since they allow us to track such mental states over time.

Summary of the analysis

For each of the three periods, we perform the following analysis:

- As mentioned above, we scale each variable so that its mean is zero and its variance equals one.
- We determine whether each variable is stationary or trend-stationary via the Augmented Dickey-Fuller (ADF) test.
- We check for collinearity among the variables via the Variance Inflation Factor (VIF).
  - We perform the dynamic vector auto-regression (VAR) analysis. The definition of VAR is in Materials and Methods and the VAR analysis implementation details are in the Supplementary Text in S1 File. This analysis consists of the following steps:• To produce the stationary versions of all the variables, we remove the trends from the trend-stationary variables and difference *action* in Period 1, where it is neither stationary nor trend-stationary.
  - We then fit the variables' stationary versions to VAR models.
  - We choose the lag order for the VAR models according to the models' information criteria and final prediction errors.

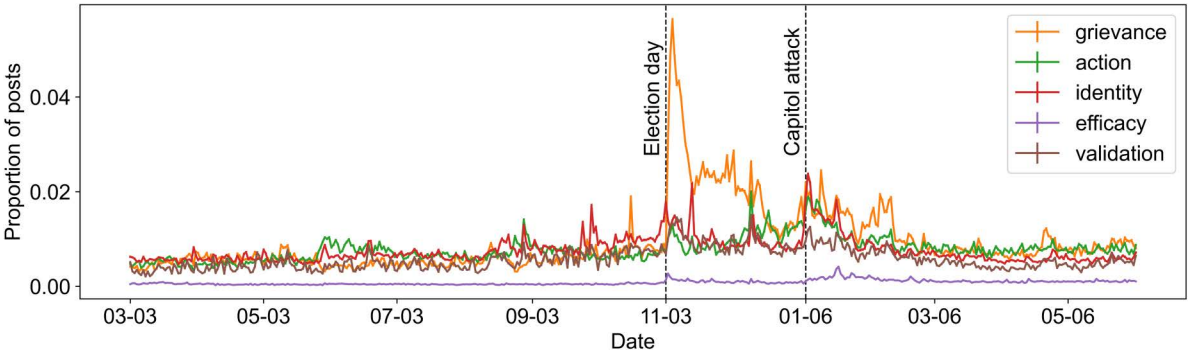


Fig 1. Daily proportion of posts around the 2020 election, versus time. Different colors correspond to different expression categories. The election day is November 3, 2020; it is marked as '11-03'. The day of the Capitol attacks is January 6, 2021; it is marked as '01-06'.

https://doi.org/10.1371/journal.pcsy.0000107.g001

- We identify the significant VAR model coefficients, which correspond to predictive interactions among the variables.
- Toward a more detailed understanding of the interactions, we also calculate impulse response functions (IRF).

In addition, we perform a pathway analysis for Period 2, to test for long-term correlations across online communities during this period. See Supplementary Text in S1 File for all details of implementation.

Trends

Table 1 indicates the status of each variable with respect to trends and stationarity during each time period. In the cases where a variable is trend-stationary, this panel specifies whether the trend is increasing or decreasing.

In order to better visualize the long-term trends in the trend-stationary and non-stationary time series, we smooth these time series by taking averages over 14-day rolling windows (see Fig 2).

In Period 1, there is a gradual rise of *identity* and *validation*, reflecting a steady growth of shared group identity and mutual reinforcement that prepare the ground for eventual actions. The dynamic of *action* shows two peaks that reflect specific mobilizing events. We hypothesize that the first peak reflects several events at the end of May 2020, including the George Floyd protests and President Trump's influential warnings that mail-in voting would lead to election fraud. The second peak could be due to a combination of several events in mid-August, including the Democratic and Republican National Conventions, as well as additional spikes in the controversy around mail-in voting.

In Period 2, the declining trend in *grievance* might indicate a shift from expressing discontent to acting on it. Once grievance about election fraud is widely recognized, the focus shifts toward actionable responses. The corresponding increase in *action* could indicate an intensified mobilization leading up to the January 6 attacks. It underscores how accumulating grievance transforms into tangible calls for mobilization. This possible causal interpretation aligns with the results of the pathway analysis (see below).

In Period 3, fast quadratic declines in *grievance*, *efficacy* and *validation* may be due to a combination of disillusionment, fear of consequences, increased scrutiny, and legal repercussions. *Identity* is strongly correlated with the other variables during this period. In a multiple linear regression model with *identity* as the dependent variable and the remaining variables as independent, $R^2 = 0.91$. This suggests that the social identity has fully consolidated by this period, so that the *identity* variable is deeply embedded in the collective narrative and is no longer independently fluctuating.

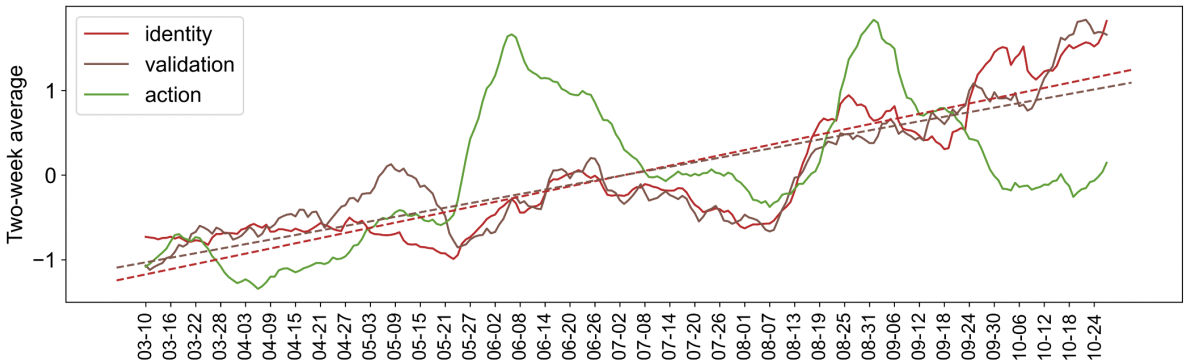
Existing theories of social identity frequently posit a decisive influence of external factors on the development of identities that underlie collective actions. Here, we have split the time series into three periods, in accordance with the key external events that we expect to fundamentally affect the development of our focal identity. This allows us to operationalize the influence of these events by considering the way that they might affect the general time series trends. Additionally, the theories frequently speak of the social identity's formation. We operationalize this latter idea as an increase in the correlation among our expression categories.

Table 1. The changing stationarity properties in the 2020/2021 dataset.

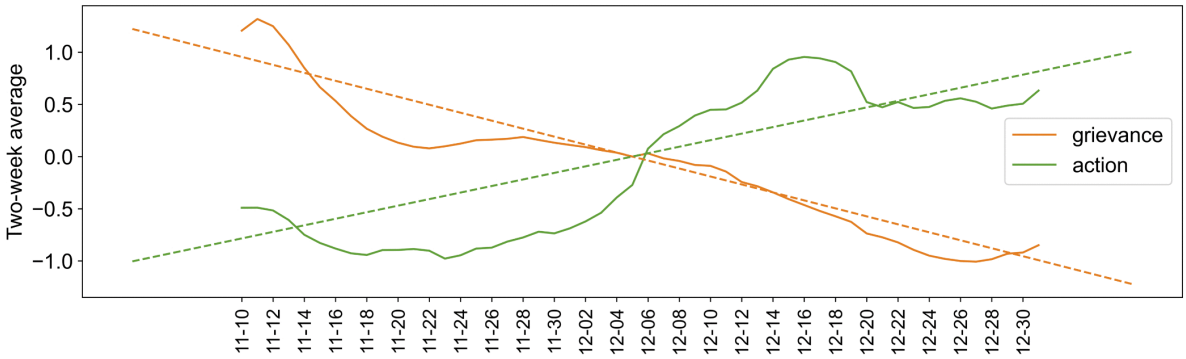
	Grievance	Efficacy	Validation	Identity	Action
Period 1	S	S	ITS	ITS	NS
Period 2	DTS	S	S	S	ITS
Period 3	DTS	DTS	DTS	S	S

S – stationary, NS – non-stationary, ITS – stationary after subtracting an increasing trend, DTS – stationary after subtracting a decreasing trend.

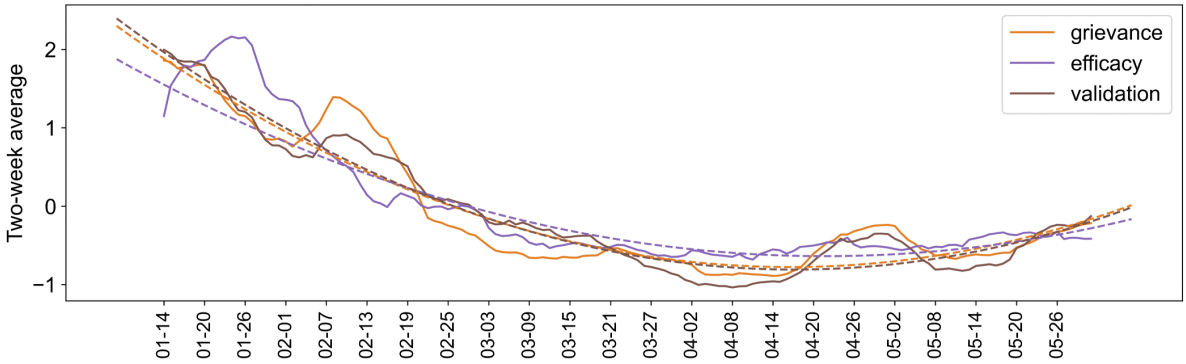
https://doi.org/10.1371/journal.pcsy.0000107.t001



(a) Period 1



(b) Period 2



(c) Period 3

Fig 2. (a) Period 1 (b) Period 2 (c) Period 3 Rolling averages of scaled post proportions for the trend-stationary and non-stationary series in the 2020/2021 dataset. Each point in a graph is labeled with the date that is at the center of the rolling window over which the average is calculated. Dashed lines indicate the regression models that we find significant; we subtract these models to obtain stationary, de-trended variables.

https://doi.org/10.1371/journal.pcsy.0000107.g002

Overall, the time series trends support a coherent narrative where online collective identity and grievance initially fuel calls for action, and the subsequent mobilization efforts culminate dramatically on January 6. Following the latter event, participants either withdraw or entrench themselves more narrowly and cautiously within their communities.

Vector auto-regression

Here, we enter the stationary, non-collinear variants of the scaled variables (or, simply, the stationary variables – see Materials and Methods) into vector auto-regression (VAR) models in order to detect significant day-to-day interactions.

Fig 3 presents the self-effects and cross-effects that we infer from the fitted VAR models, in the form of the regression coefficients that are significant at the 5% level. These effects are the significant predictive interactions among the stationary portions of the scaled variables (and the differenced *action* in Period 1) on daily timescales. Impulse response functions in S2, S3 and S4 Figs show that the effects propagate over 3–4 day horizons. In the rest of this section, we provide an interpretation of Fig 3 as predictive associations consistent with prior theory.

The self-lag coefficients (self-lags) of the stationary and detrended variables are comparable in magnitude and positive, except for differenced *action* in Period 1. Positive coefficients reflect each variable's daily persistence or inertia. The negative coefficient for differenced *action* indicates that spikes in discussions about actions are followed by declines in discussion intensity. During Period 3, *action* exhibits the highest self-lag among all variables across periods, indicating a strong stabilization of this type of expression by this time – as well as stabilization of *identity*, which is highly correlated with *action* during this period ($r = 0.92$, also see the corresponding variance inflation factors and regression results in Materials and Methods).

We interpret the cross-lag coefficients (cross-lags, with values in parentheses) during Period 1 as follows. Grievance expression reliably predicts validation (0.14), meaning that airing grievances anticipates group affirmation, reinforcing their legitimacy. Validation promotes identity formation (0.15), as affirmation from peers solidifies collective identity. Expressing common identity strongly bolsters efficacy (0.32), increasing belief in the success of collective action. Efficacy maintains grievance expression (0.11), suggesting that higher efficacy sustains the airing of grievances.

Thus, Period 1 reveals a clear developmental sequence: *grievance* → *validation* → *identity* → *efficacy* → back to *grievance*, confirming psychological theories of social identity formation. This loop-like structure supports theories where group validation and identity formation reinforce each other, creating stable dynamics [30,40].

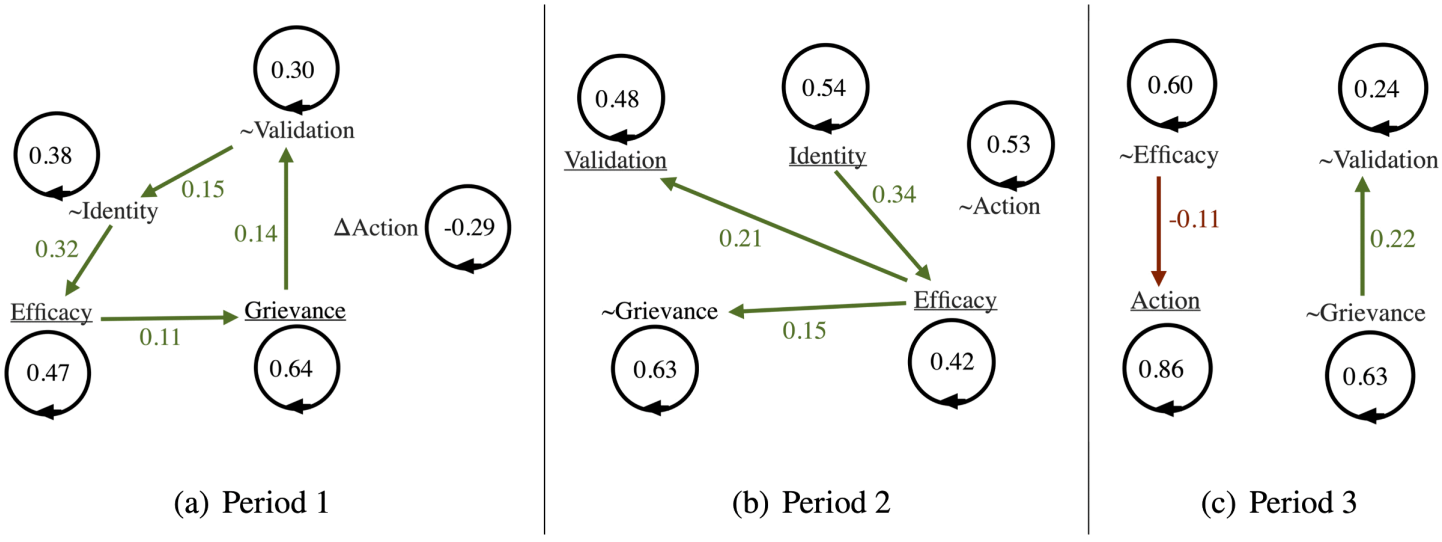


Fig 3. (a) Period 1 (b) Period 2 (c) Period 3 Self-effects and cross-effects in the VAR models for the 2020/2021 data. The effects take the form of coefficients that are significant at the 5% level. The names of the stationary variables are underlined, those of the detrended variables are preceded by tildes, and the name of the differenced variable is preceded by Δ. The self-lag coefficients are inside circles, the cross-lag coefficients are next to straight arrows. For each coefficient, the lag is one day.

https://doi.org/10.1371/journal.pcsy.0000107.g003

Our interpretation of the Period 2 cross-lags is as follows. Social identity strongly shapes beliefs about group efficacy (0.34). Efficacy expression increases grievance expression (0.15), suggesting that confidence in group power sustains the relevance of grievance. Efficacy also boosts validation (0.21), indicating that belief in group efficacy encourages mutual affirmation among members.

In sum, during this critical rallying phase, the collective identity strongly shapes collective confidence (i.e., efficacy beliefs), which then reinforces both grievance expression and validation. This sequence closely matches the social identity model of collective action [30], demonstrating that identity-driven efficacy is central to sustaining group cohesion and grievance expression. *Action* is not predicted by the other variables.

In Period 3, much like in Period 1, expression of grievance reliably prompts validation (0.22), meaning mutual affirmation remains important even after the mobilization peak. Increased efficacy in Period 3 reduces explicit calls for immediate action (−0.11), suggesting that confidence in group power may lower action urgency – potentially indicating a consolidation phase where efficacy beliefs persist without the accompanying calls to action.

Identity is not present explicitly in Period 3 due to its extremely high collinearity with the other variables (i.e., it is highly predictable by the other variables via linear regression; see above).

Overall, Period 3 marks a consolidation of the social identity: a strong, stable internal persistence across the variables, sustained validation of the grievance, and a notable reduction in immediate calls for action with increased efficacy. *Identity*, fully consolidated, no longer varies independently. This supports the idea of social identity becoming deeply internalized and self-reinforcing, which is consistent with the long-term stabilization that is predicted by theory [40,44].

On the whole, the VAR results reveal a predictive sequence that supports the following developmental trajectory for the sub-identity unified by shared beliefs in election fraud. This sub-identity emerges within the MAGA movement through a robust cycle from *grievance* through *validation* and *identity* expression toward *efficacy*. The identity then mobilizes its followers during Period 2, shaping the collective efficacy belief, as well as sustaining the expressions of grievance and validation. During this period, high daily self-lag coefficients for all five expression categories suggest strong collective readiness; for example, the amount of *grievance* expression on any given day strongly predicts the amount of this expression on the following day. Finally, the identity consolidates in Period 3, with the *identity* variable fully predicted by the other variables, grievance expression that continues to drive validation, and a robust equilibrium indicated by strong internal self-lags. Increased efficacy now reduces immediate calls for action, suggesting a post-event settling.

These findings align with those of Smith, van Zomeren and coauthors [30,40]. In particular, the sequence of *grievance*→*validation*→*identity*→*efficacy*→*action* is explicitly confirmed, consistent with the canonical developmental pathways in the literature. Here, mutual validation emerges explicitly as a critical factor stabilizing collective identity and grievance expression, aligning with the emphasis on validation in the work of Smith and her colleagues [40].

Furthermore, the inverse relationship between efficacy and explicit action calls in Period 3 provides an intriguing extension to the current theories: increased confidence in collective efficacy may reduce the immediate need or urgency for collective mobilization after major collective events, reflecting a psychological shift toward identity consolidation rather than continuing action.

Pathway analysis

As shown above, statistical analysis of the daily time scales via vector auto-regression reveals the predictive pathway *grievance*→*validation*→*identity*→*efficacy* on this time scale during Period 1. This pathway is a close match to the theory due to Smith and coauthors [40]. On longer time scales, we have identified increasing trends in *validation* and *identity* in Period 1, as well as decreasing trends in *grievance*, *efficacy* and *validation* in Period 3. We hypothesize that these trends are due to exogenous factors such as the impending election in Period 1 and the receding of the attacks into the past during Period 3. Additionally, according to our analysis of collinearity (see Materials and Methods), all five variables are

strongly collinear in Period 3. Overall, our analysis of Periods 1 and 3 suggests the absence of long-term predictive interactions among the variables during these two periods.

The variables' behavior is markedly different during Period 2. In particular, this period exhibits a decreasing trend in *grievance* and an increasing trend in *action*. In combination with the fact that this is the period between the factual or alleged realization of the grievance and the occurrence of the corrective action, these trends suggest the possibility that they are not entirely due to exogenous factors and that there is a causative relationship between them. Such a suggestion gains additional support when we examine the middle panel of S1 Fig and notice that, apart from the synchronous peak in all variables on December 12th, the temporal sequence of the highest points in four of the variables, *grievance*→*validation*→*identity*→*action*, closely follows the above-mentioned theoretical pathway from the literature.

In accordance with this qualitative observation, we conduct a pathway analysis, akin to that in a recent study by Smith and colleagues [3]. In particular, we sample across communities in the social media universe, similarly to the way Smith and coauthors [3] sample across individual social media users. We also distinguish among four non-intersecting time intervals in Period 2. Time 1 is an interval around the *grievance* peak at 11/06. Time 2 – an interval around the *validation* plateau between 11/08 and 11/10. Time 3 is an interval around the *identity* peak at 11/15. Time 4 is between 12/17 and 01/04, when *action* is prominently higher than the other variables.

For each community, we determine the number of *grievance* posts at time 1, the number of *validation* posts at time 2, the number of *identity* posts at time 3, and the number of *action* posts at time 4. The pathway analysis is a series of multiple regressions across the communities, where all the variables that precede a dependent variable in time are included as independent. In other words, we regress *validation* on *grievance*, *identity* on *grievance* and *validation*, and *action* on the other three. We assume that the residuals are normal, in accordance with common practice. To make sure our results do not depend on the exact boundaries between the four time intervals, we run the analysis with five different sets of intervals and compare the results (see Supplementary Text in S1 File). Fig 4 presents a summary of all the results; this figure is equivalent to Fig 2 in the article by Smith and colleagues [3]. Here, the arrows between different variables are predetermined by the analysis design, whereas the significances, the signs and the magnitudes of the corresponding coefficients are due to the data. The coefficients that are predicted by Smith and coauthors [40] are positive and largely significant, offering support for their theory on the time scale of the entire second period. In addition to the predicted effects, we find a direct effect of *grievance* talk on *identity* expression and a direct effect of *validation* on calls to *action*.

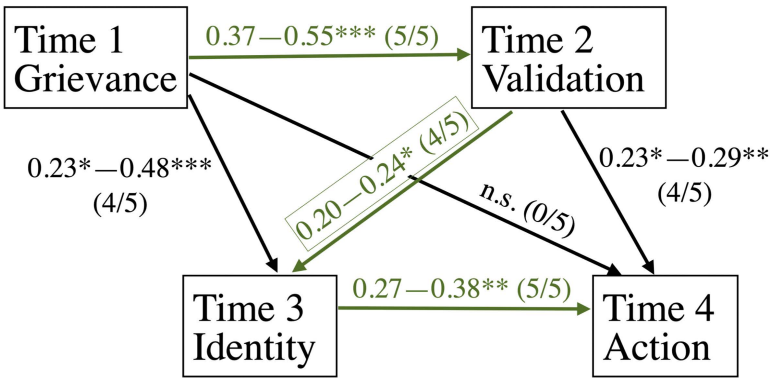


Fig 4. Results of the pathway analysis for Period 2 communities in the 2020/2021 data. Five different sets of boundaries between the time intervals were used. The coefficients that are predicted by the psychological theory are in green. N.s.: not significant ($P \geq 0.05$), one star: $P < 0.05$, two stars: $P < 0.01$, three stars: $P < 0.001$. For example, the text above the *validation*→*identity* arrow means that 4 out of 5 boundary sets produce significant coefficients, with a magnitude range between 0.20 and 0.24 and $P < 0.05$. The excluded community, 4Chan1, has thousands to tens of thousands of posts in each category.

https://doi.org/10.1371/journal.pcsy.0000107.g004

Social identity theories are based on the idea of logical connections between the collective mental states that reinforce each other with time. The above VAR and pathway analyses operationalize this idea in our time series, by seeing whether the expression in one of the categories at a given time reliably predicts expression in another category at a later time. In accordance with the above results, the signal that we detect generally aligns well with the existing theories, although the results also suggest theoretical refinements and additions.

2024 elections

We now apply our analytic methodology to the "hate universe" data around the 2024 presidential election, then compare and contrast our results between 2020 and 2024.

Time series. The 2024 election occurred on November 5, 2024 (i.e., 11/05/24). To track the development of the focal social identity around this event, we analyze ~30 million posts between 02/01/24 and 03/01/25. Fig 5 shows the frequencies of posts in different categories during this period of time. Here, the regular expressions for each expression category are the same as in our main analysis. Due to our dataset's technical limitations (see Materials and Methods), we split the pre-election period in two, thus conducting our analysis separately on three periods: Period 1a, which is before the first grayed out interval in Fig 5; Period 1b, which is after this interval and before the election; and Period 2, which is after the election and includes the second grayed out interval. As in our analysis of the 2020 elections data, we scale the time series so that each has zero mean and unit variance in any given period. The scaled time series are shown in S6 Fig.

We verified using an LLM that our regular expression filters capture very similar types of grievance and action between 2020 and 2024 (see Supplementary Text in S1 File). Apparently, for example, the grievance with regard to election fraud and the illegitimacy of the Biden administration remain topical and the calls for violent collective action remain popular in 2024. Nevertheless, as we show further in this section, the interactions among the different expression categories change dramatically between 2020 and 2024.

Trends. Fig 6 illustrates the observed trends via 14-day rolling averages. In Period 1a, there is a statistically significant linear increase in *identity* ($P < 0.05$ for the ADF stationarity test with linear trend), accompanied by a quadratic increase in *grievance*, although the latter is significant only at the $P = 0.1$ level. A number of major events external to the social media discussion could be the causes of several local peaks in these variables. In particular, *identity* has a considerable peak around Donald Trump's conviction in *The People of the State of New York v. Donald J. Trump* on May 30. Additionally,

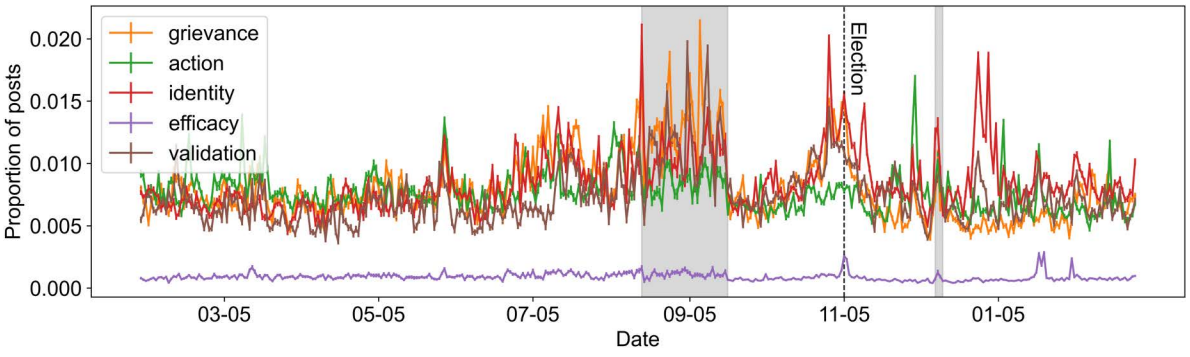
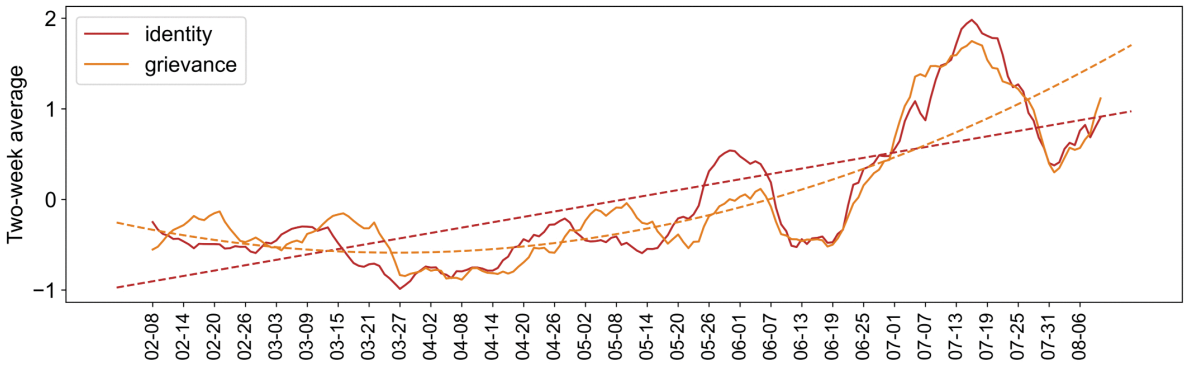
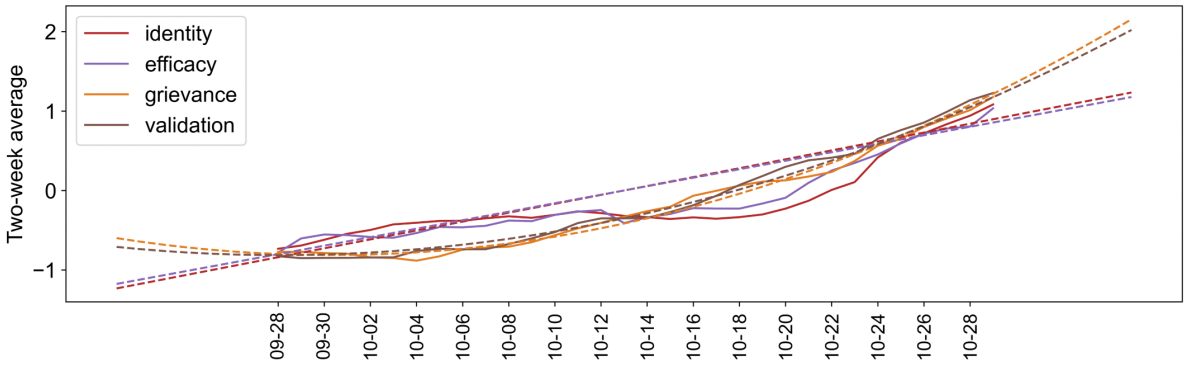


Fig 5. Daily proportions of posts versus time, around the 2024 presidential election. Different colors correspond to different expression categories. The election day is November 5, 2024; it is marked as '11-05'. The grayed out intervals are between August 17 and September 20 of 2024 and between December 11 and December 14 of 2024; during these intervals, only partial data is available due to instabilities in the 4Chan social media platform, which provides the majority of the data. Period 1a is up to the beginning of the first grayed out interval, Period 1b begins with the interval's end and concludes with the election, Period 2 starts with the election and includes the second grayed out interval.

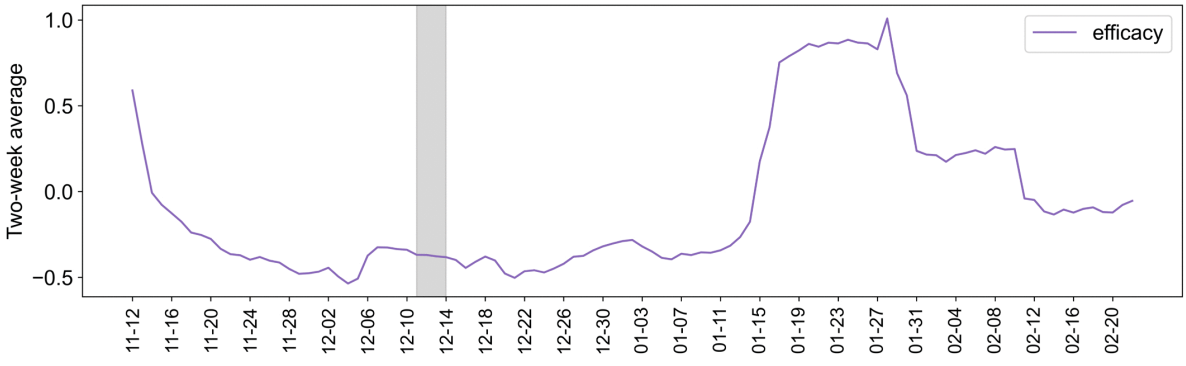
https://doi.org/10.1371/journal.pcsy.0000107.g005



(a) Period 1a



(b) Period 1b



(c) Period 2

Fig 6. (a) Period 1a (b) Period 1b (c) Period 2 Rolling averages of scaled post proportions for the trend-stationary and non-stationary series in the 2024/2025 dataset. Stationarity is determined according to the Augmented Dickey-Fuller (ADF) test and, for the most part, the $P<0.05$ significance criterion. Each point in a graph is labeled with the date that is at the center of the rolling window over which the average is calculated. Straight dashed lines indicate the linear regression models that we find significant; we subtract these models to obtain stationary, de-trended variables. Curved dashed lines indicate quadratic trends that the ADF test finds significant at the $P<0.10$ but not the $P<0.05$ level; we do not detrend the corresponding variables. Only partial data is available during the grayed out interval.

https://doi.org/10.1371/journal.pcsy.0000107.g006

both *identity* and *grievance* show large peaks between the Biden-Trump debate on June 27 and Joe Biden's withdrawal from the electoral race on July 21. The remaining variables are stationary during this period.

During Period 1b, we observe statistically significant linear increases in *identity* and *efficacy* ($P < 0.05$), while quadratic increases in *grievance* and *validation* are present but significant only at $P = 0.1$. *Action* is stationary during this period. In Period 2, all variables except *efficacy* become stationary. Notably, around the Inauguration Day on January 20, *efficacy*'s running average exhibits values that are noticeably higher than during the rest of this period.

VAR. Due to technical limitations of our data, we do not conduct a VAR analysis of Period 1b (see Materials and Methods). The results of the Period 1a and Period 2 VAR analyses are in Fig 7.

During Period 1a, *action* (self-lag coefficient equal to 0.59), *validation* (0.51), and *identity* (0.49) exhibit strong persistence, indicating stable daily continuity within these categories. Differenced *grievance* has a negative self-effect (-0.24), indicating that expressed grievances on a given day likely decline on the next day, possibly due to a fluctuating discussion focus.

During Period 2, level *grievance* (0.70) is rather self-persistent, suggesting a day-to-day discussion of the focal grievance that is more consistent than in Period 1a. The self-lag coefficients for *identity* (0.58) and *validation* (0.47) are just as high during the post-election Period 2 as they are during the pre-election Period 1a, indicating that the election and its results do not have an effect on these variables' daily continuity. Differenced *efficacy* shows strongly negative persistence (-0.42) during Period 2, indicating considerable volatility or fluctuations in collective efficacy beliefs post-election.

The negative effect of *identity* on *grievance* (cross-lag coefficient equal to -0.14) in Period 1a suggests that expression of belonging to the collective identity reduces the inclination to express grievance – a marked deviation from 2020, where this kind of grievance expression is foundational to identity formation. However, *validation* still drives *action* (0.15), maintaining a limited but important role in promoting the social identity's normative action. Overall, unlike the 2020 pre-election period, no strong predictive pathways from grievance to action through identity and validation emerge. This suggests that activity related to the focal identity was driven primarily by internal persistence in 2024, rather than by external grievances, as in 2020.

In Period 2, *identity* significantly reinforces *validation* (0.19), highlighting a collective identity that fosters mutual affirmation. Increased *action* in this period slightly reduces *grievance* (-0.12), perhaps suggesting that once the need for mobilization is articulated, grievances may seem redundant or less immediate. Expressing these grievances strongly decreases

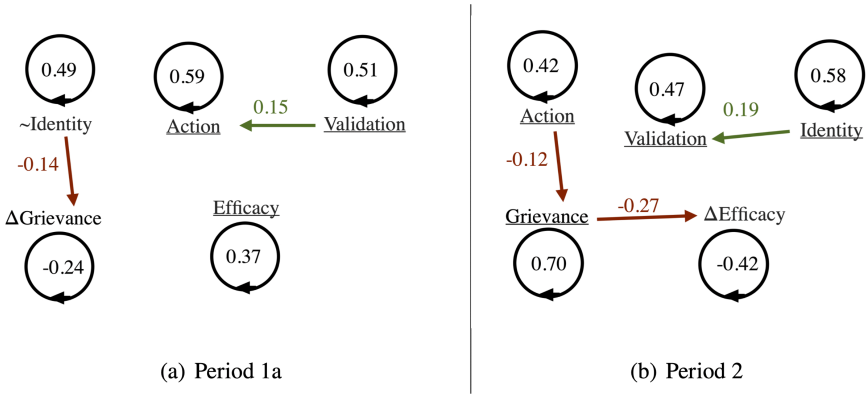


Fig 7. (a) Period 1a (b) Period 2 VAR analysis of the 2024 data. Self-effects and cross-effects inferred from the fitted models, in the form of the coefficients that are significant at the 5% level. Names of the stationary variables are underlined, those of the detrended variables are preceded by tildes, those of the differenced variables are preceded by Δ. Self-lag coefficients are inside circles, cross-lag coefficients are next to straight arrows. For each coefficient, the lag is one day.

https://doi.org/10.1371/journal.pcsy.0000107.g007

collective efficacy (-0.27), potentially indicating that these grievances are no longer associated with the normative action, the way they are in 2020 and 2021.

To elaborate further, the periods immediately following the presidential elections show marked differences between 2020 and 2024. While the focal identity and the validation process still exist in 2024, calls to action no longer clearly stem from efficacy beliefs or validation processes, and the negative relationship between *grievance* and *efficacy* during this election cycle suggests a reduced group coherence. The peak in *efficacy* around the Inauguration Day suggests an increase in collective optimism or perceived group strength, likely influenced by the inauguration event itself.

Summary. Overall, although the analyses of the 2020 and 2024 presidential elections yield some similarities in the social identity processes between the two elections, there are many important differences as well. These differences reflect significant political and psychological contrasts between the two election cycles.

In particular, during 2020 and 2021, the perceived injustice of losing the election due to fraud appears to trigger a structured development of collective identity, moving systematically from grievance expression through validation and identity formation towards efficacy and mobilization. On the other hand, in 2024 and 2025, Trump's electoral victory does not provide a grievance of comparable clarity or strength. Consequently, expression of grievance is negatively associated with the expression of belonging to the focal identity before the election, indicating fundamentally different psychological dynamics, in which early identity expression replaces, rather than emerges from, grievance articulation.

In 2020, validation is a critical driver of identity formation and mobilization. The trajectory is clear – validation forecasts the expressions of belonging to the focal identity and of belief in the normative action's efficacy, enabling sustained mobilization for action. In contrast, in 2024, validation positively impacts calls for action early on (in Period 1a), but overall validation and action remain less dynamically integrated into a coherent narrative. Validation persists but does not catalyze sustained calls for collective mobilization, and calls to action are less systematically connected to other identity elements.

In 2020, expressed beliefs of efficacy consistently play a motivational and sustaining role, reinforcing grievance and validation expressions and directly contributing to collective mobilization. On the other hand, in 2024, efficacy beliefs show considerable volatility and instability, particularly post-election. Grievance expression negatively impacts efficacy beliefs, suggesting that such expression is no longer associated with the corrective actions that are discussed in 2020.

The differences between the processes in 2020 and 2024 are consistent with the psychological theories of social identity [30,40,44], confirming the pivotal role of external factors and emphasizing the necessity of a clear grievance or a perceived collective injustice to activate robust identity and mobilization dynamics. The absence of such a clear grievance in 2024 appears to reduce the expression of the focal identity, weaken the efficacy dynamics, and decrease the calls for mobilization.

The 2024 results suggest that while social identity processes still maintain stable categories of expression such as *identity* and *action*, the lack of a clearly shared collective grievance and the volatility of the belief in the corresponding normative action's efficacy significantly reduce the potential for the sort of collective mobilization that occurred in 2021. This underscores that the expressions of grievance and perceived injustice are foundational for collective mobilization through identity formation processes.

Discussion

This article operationalizes key components of social identity and collective action theories in large-scale social media discourse, then evaluates where our data confirms the established accounts, where the accounts require refinement, and where they break down. We focus on five recurring components that we operationalize as types of expression – grievance (a perceived injustice), validation (affirmation and consensus-building), identity (a shared in-group identification), efficacy (beliefs about collective capacity), and action (normative calls for what should be done) – and on the dynamic interactions among these components in collective communication. These interactions can require both logical reasoning and emotional framing: for example, a shared grievance can support the conclusion that a corrective action is necessary

[40]; simultaneously, the same grievance claim can contribute to identity formation only when expressed within a specific emotional framework [2,31]. We qualitatively observe the requirement for such frameworks in our data, even though we do not quantify it.

To trace the above-mentioned dynamics around the 2020 and 2024 U.S. presidential elections, we combine trend inspection, stationarity tests, vector autoregression (VAR), and pathway analysis. These tools allow us to assess (i) which theoretical components or types of expression remain stable or shift sharply across event-defined periods and (ii) whether one component predicts later levels of another, holding other components constant. The present Discussion section proceeds in three steps: we first contrast the overall results across the election cycles, then interpret each type of expression in turn, and finally summarize what our results imply for theory and methodology.

Our analyses are observational. VAR and pathway analysis describe temporal associations, not causal effects; they characterize aggregate discourse dynamics within selected communities and time windows, with substantially weaker support for causal claims than evidence from controlled experiments. Furthermore, our design cannot isolate the influence of any specific leader or speech act. Accordingly, our findings should not be read as attributing motives, coordination, or wrongdoing to President Trump or to any other individual.

2020 versus 2024

We split each election cycle into event-defined periods. For 2020, we analyze the pre-election period (Period 1), the post-election period through January 6 (Period 2), and the post-January 6 period (Period 3). For 2024, we analyze two pre-election sub-periods (Periods 1a and 1b) and a post-election period (Period 2). We find both similarities and differences between the cycles.

In the case of the 2020 cycle, the results are broadly consistent with the canonical models of collective action, in which grievance predicts validation, validation predicts identity, and identity predicts efficacy and action [2,30,31,40]. This canonical sequence is particularly salient during Periods 1 and 2; the combined evidence for all three periods suggests a progressive alignment among the theory components that culminates in a coordinated discourse around action by January 6.

The results for the 2024 cycle do not reliably support this sequence. Instead of a mutual reinforcement among the expression types, these are often weakly or negatively related. Grievance can be associated with weaker efficacy, action discourse can predict decreases in grievance, and identity predicts validation without consistently predicting efficacy beliefs or action talk. Rather than exhibiting a progressive alignment, the 2024 cycle presents a more fragmented configuration, suggesting that the mobilizing role of a focal grievance is context-dependent and may weaken when it is not widely validated or is inconsistent with the currently relevant political outcomes.

Operationalized theory components

Below, we connect our results to the established theories of collective action, focusing on the Social Identity Model of Collective Action (SIMCA) [30], the normative alignment model [31], and the identity-norm nexus (INN) [2,40]. We organize this discussion around the five theoretical components that we operationalize – grievance, validation, identity, action, and efficacy – and emphasize how the coupling that we observe between these components differs between 2020/2021 and 2024/2025.

Grievance. Many social-psychological theories of collective action treat the perceptions of grievance and injustice as foundational. SIMCA frames the injustice appraisals as one of the three antecedents to mobilization, and related accounts emphasize the pivotal role of the perceived gaps between the actual societal conditions and those that would be considered equitable [30,45]. Our 2020 results are consistent with this functional importance of grievance perception. Grievance expression regarding alleged voter fraud is relatively stable before the election and spikes immediately after the election. Across the three periods, VAR and regression analyses show grievance predicting increases in validation and identity

expression. Here, the fraud claims appear sufficiently plausible within the studied communities to elicit agreement and to support the build-up of an associated in-group narrative.

In 2024, the grievance component behaves differently. Grievance expression does not consistently lead to validation or identity talk, and in the post-election period it can negatively predict efficacy while action discourse negatively predicts grievance. In other words, grievance is sometimes associated with a weakening, rather than a strengthening, of other mobilizing components. A plausible interpretation is contextual: in 2024, the Republican candidate both appears viable pre-election and wins the election. This likely reduces the perceived plausibility of fraud claims, relative to 2020, and weakens the capacity of grievance to generate validation and identity. More generally, these results suggest that grievance is not a uniform driver; its mobilizing potential depends on whether it is aligned with the perceived reality.

Validation. Validation corresponds to affirmation and agreement, and potentially leads to the emergence of shared norms through interaction. Accounts of inductive identity formation emphasize that consensus-building can be central to the emergence of a shared identity in discussion [28,29,46,47]. In our 2020 results, validation plays a bridging role. Validation is tightly linked to grievance and identity, and pathway analysis shows validation predicting identity and action in the period between the election and January 6. In this period, validation appears to function as the hinge by which grievance claims are transformed into a shared identity and coordinated action discourse.

In 2024, validation's role shifts. In one pre-election sub-period, validation predicts action, suggesting that agreement can still energize mobilization. However, in the post-election period, validation is predicted by identity rather than by grievance, indicating a reversal of the expected sequence. In this configuration, validation may reflect the already established in-group positions, instead of producing a consensus around a newly salient injustice. This pattern reinforces the centrality of validation in interactive models, while showing that the direction of coupling can vary with context.

Identity. Social Identity Theory and Self-Categorization Theory describe identity as a basis for collective behavior [24,25]. The INN account further emphasizes a co-emergence of identity and action norms through discussion [2,40]. Our 2020 results support these interactive accounts. Identity expression increases across the pre-election period, is predicted by validation, and predicts efficacy beliefs and action talk. During the post-election period leading to January 6, identity is linked to action talk in pathway analysis, consistent with identity as a proximate motivator of collective action. In this cycle, identity is both a product of validation and a driver of efficacy and action discourse, consistent with a convergence on a corrective action norm under a perceived threat.

In 2024, identity remains active but functions differently. In the first pre-election sub-period, identity negatively predicts grievance, suggesting a decoupling from the perceptions of injustice. In the post-election period, identity predicts validation but not action. Thus, identity discourse can persist without aligning grievances and action. A contextual interpretation is again plausible: if the fraud claims are less credible, the perception of threat may be insufficient to produce a strong identity growth around the focal grievance or to translate identity into action norms. These results highlight a variability in identity's role: identity can bridge grievance and action in some contexts (2020/2021), but in others (2024/2025) it may not produce a mobilizing alignment.

Action. Action discourse refers to the expressions of what should be done, ranging from calls to protest to demands for institutional change. Normative alignment theory emphasizes that action emerges when grievance, identity, and efficacy are aligned within a shared normative frame [30,31]. In 2020/2021, action discourse rises steadily between the election and January 6, and pathway analysis shows validation and identity predicting action in the run-up to the attack. These patterns support the view that action talk crystallizes when grievances are validated and linked to identity.

After January 6, we observe that efficacy can negatively predict action in VAR. One interpretation is that an increase in the perceived feasibility of collective action such as the Capitol attack due to the success of the attack, in combination with the attack's observed consequences (including deaths and legal risk), reduces the subsequent advocacy for similar collective action, shifting the discourse away from mobilization.

In 2024/2025, action discourse is less connected to other expressions. In one pre-election sub-period, validation predicts action, but in the post-election period action negatively predicts grievance. Thus, rather than being a culmination of consensus around a focal injustice, action talk can sometimes undermine or displace grievance expression. This challenges a simple view of action as an endpoint of alignment and suggests that action discourse can operate independently of, or even in tension with, the central grievance narrative.

Efficacy. In SIMCA and related theories, efficacy beliefs mediate the relationship between identity and action [30,31]. Our 2020/2021 results show efficacy embedded in a broader dynamic system: identity predicts efficacy, efficacy is linked to grievance and validation, and efficacy can negatively predict action after January 6. These patterns suggest that efficacy is not a stand-alone belief but a discourse component shaped by identity that, in turn, shapes how grievances and action are discussed. The post-January 6 negative relationship between efficacy and action may reflect a decreased mobilization advocacy after a major collective event, consistent with the above-mentioned learning about both the feasibility and the costs of such events.

In 2024/2025, efficacy is both more fragile and less connected to other expressions. In the post-election period, grievance negatively predicts efficacy, suggesting that injustice expression can weaken the beliefs in collective capacity rather than strengthening them. This again underscores context sensitivity: grievances may mobilize only when they are validated and aligned with the identity and action norms; otherwise they may erode the collective confidence or reflect a resignation rather than the need for mobilization.

Confirmation, extension and refinement of theory

Taken together, the 2020/2021 results confirm several interactive accounts of collective action. Across multiple analyses, grievance predicts validation, validation predicts identity, and identity forecasts efficacy and action, consistent with SIMCA, the normative alignment theory, and the INN model [2,30,31,40]. In this cycle, validation appears central for translating grievance into shared identity and action norms, and efficacy serves as an important intermediary.

By contrast, the 2024/2025 results refine these accounts by showing that the canonical sequence can fail. The absence of consistent grievance→validation→identity→action coupling, alongside several negative relationships, suggests fragmentation rather than alignment. A practical implication is that models of mobilization may need to treat grievance validation and identity-action alignment as contingent on the context and the event structure rather than as the default. In particular, when a focal grievance is not credible or socially reinforced, the same discourse components may coexist without producing a shared pathway toward action.

The contrast between the two cycles also highlights the interplay of top-down events and bottom-up interaction. In 2020/2021, top-down events (the election outcome and January 6) coincide with sharp shifts in grievance expression and with bottom-up reinforcement via validation and identity discourse, consistent with accounts in which elite cues and salient events interact with grassroots consensus-building [1,38]. In 2024/2025, comparable top-down triggers that would reinforce grievance and align it with identity and action are weaker or absent, and bottom-up reinforcement is correspondingly weaker, producing a more fragmented system.

Methodologically, this study demonstrates that stationarity testing, VAR, and pathway analysis can be used to operationalize theory components in large-scale communication data and to evaluate theory-linked temporal predictions in naturalistic settings. In the same vein as prior digital-trace work on protest participation [34], we map theoretical constructs onto discourse, providing a replicable framework for testing and refining psychological theory with observational data.

Both theoretical work and empirical studies suggest that radicalization and mobilization can exhibit cascade-like dynamics and self-organized criticality (SOC) [13,18,48–50]. Our analysis does not show any signatures of SOC. In particular, we do not see any clear cascade-like precursors immediately before January 6, and the sharp post-election grievance increase is consistent with an external shock rather than a self-reinforcing communication cascade. Nevertheless, the alignment dynamics we observe in 2020/2021 may be foundational for the radicalization and offline mobilization

toward the Capitol attack, particularly if additional processes (including higher-frequency cascades, network diffusion, or threshold dynamics) are present in this setting. A dedicated SOC analysis would require measures tailored to cascade processes (e.g., cascade-size distributions, branching-style reproduction metrics, or critical-transition diagnostics) and finer temporal resolution than our daily aggregates.

Qualitative features of the "hate universe" data, alongside our formal analysis, connect our results to broader accounts of online activism while also highlighting important differences between these accounts and our work. Firstly, our data suggest that salient events (e.g., election fraud claims or high-profile court cases) drive several of the largest discourse spikes, consistent with the idea that topic interest can be a strong predictor of participation [51]. Secondly, users frequently frame moderation or fact-checking as an attack on the movement, which can increase posting volume; this echoes broader "backfire" accounts in which perceived repression galvanizes engagement [52,53]. Thirdly, we observe substantial toxic out-group language, which appears to generate in-group engagement while discouraging cross-cutting conversation, consistent with evidence that out-group interactions can be more toxic and less engaging [54]. Fourth, the networks that we study include newcomers who may encounter approachable in-group arguments; this resonates with the work on allyship and legitimacy, where the engagement by advantaged-group allies can shape a movement's dynamics [55]. Finally, although some literatures emphasize a mix of nonviolent and violent tactics in successful mobilization, discourse in our dataset often favors violent tactics and downplays peaceful alternatives [56]. Together, these comparisons suggest that right-wing mobilization around U.S. elections exhibits distinctive rhetorical cues and strategic adaptations that merit further study.

Consistent with recent work, conspiracy-oriented networks can validate grievances and circulate violent planning cues [35], and mobilization to violence can be associated with discourse about efficacy and action logistics [36,37]. Our results complement these studies by showing that, in 2020/2021, grievance validation and identity-action alignment are jointly present in the run-up to January 6, whereas in 2024/2025 similar components can persist without producing the same developmental pathway.

Study limitations

Our results may depend on the specific nature of the "hate universe" dataset and on the collective action context that we study. In particular, the combination of a selection for extremist content and our particular platform choice may bias both the prevalence estimates and the inferred pathways. Accordingly, any generalization should proceed with care: the results speak most directly to the discourse dynamics within the communities and time windows that we study, not necessarily to social media populations at large.

The analytic approach has additional limitations. First, the analyses are observational and assess predictability rather than causal effects; causal hypotheses require confirmation through experiments or further observational designs. Second, the measurement error from regular-expression filters can generate both false positives and false negatives. We estimate that the false positives have a limited impact on the qualitative conclusions if they do not introduce an appreciable bias, but time-varying error or category-specific error can bias the regression coefficients. Sarcasm in validation expression and retrospective mentions of past action in the supposed calls for future action can make such inaccuracies difficult to detect. Third, statistical testing in VAR raises multiple-comparisons concerns. For example, with $P < 0.05$ and five variables at lag order one, there are 20 coefficients and roughly one false positive expected by chance; future work may adopt stricter thresholds or correction procedures. Finally, our use of community distinctions is limited to a subset of analyses and does not incorporate network structure or platform-level diffusion.

Our statistical choices should not be seen as fixed. We keep the methods relatively simple because the overall analytic framework is novel, but several refinements to the statistical methodology may be natural in future work. One could handle collinearity via regression regularization; future research could analyze non-stationary series with co-integration tests and error-correction terms rather than differencing alone; and Bayesian VAR could stabilize the regression coefficient

estimates, incorporate prior knowledge, and provide lag selection that is more robust. Likewise, rather than incorporating the major events via the splitting of the time series into periods, future models could incorporate exogenous time-varying covariates that represent the event effects, potentially allowing analysis over longer continuous time windows. Additional robustness checks are also important: repeating the pathway analyses under alternative normalizations (e.g., proportions rather than raw counts), varying the lag orders in VAR, and performing sensitivity analyses that inject plausible levels of measurement noise into each series would help quantify the dependence of specific regression coefficients on the modeling and preprocessing choices. Future analyses may also adopt multiple-testing adjustments (e.g., false discovery rate control) when interpreting sets of VAR coefficients. Finally, finer temporal sampling (e.g., hours rather than days) would allow for a better detection of short-lived surges, a better resolution of the responses to exogenous events, and a more sensitive identification of self-reinforcing cascades.

Future work can improve the robustness and external validity of our framework by applying it to other movements and to non-hateful speech, by incorporating additional data modalities (e.g., images and memes), and by using a finer temporal sampling to resolve faster dynamics. Accurate NLP classifiers (e.g., ones that measure emotions) and richer representations of platform and community structure (e.g., ones that include network ties, exogenous event covariates, or diffusion measures) can reduce the measurement error and better capture the mechanisms that link discourse to mobilization.

Conclusion

The 2020/2021 cycle supports the canonical interactive accounts of collective action, whereas the 2024/2025 cycle reveals context-dependent fragmentation and negative coupling among key theoretical components. These results strengthen the communicative models of collective action that emphasize social identity, mutual validation and efficacy beliefs while underscoring the limits of these models across political contexts. More broadly, the framework provides a replicable bridge between social-psychological theory and large-scale discourse analysis.

Materials and methods

Ethics statement

The George Washington University Committee on Human Research determined that our study is exempt from Institutional Review Board (IRB) review. Our data consist only of anonymous, publicly available, community-level information.

"Hate Universe"

Our analysis is based on the data from the "hate universe" – a collection of communities on several social media platforms and these communities' posts over an extended period of time. We build this dataset using the same methodology as prior publications [42,43,57,58], expanded to additional platforms. The platforms for the 2020–2021 dataset are Rumble, YouTube, Gab, 4Chan, Bitchute, VKontakte, Facebook, Telegram, Instagram, and Twitter. The platforms for the 2024–2025 dataset are the same, with the exception of Twitter and Instagram. Previously, we have utilized the 2020–2021 dataset to explore the changes in the hate universe's network topology through the 2020 presidential election and the subsequent Capitol attacks [43].

The meaning of a "community" is platform-specific. For example, on the Russian state-controlled platform VKontakte, a community is called a Club. On Facebook, a community is called a Page; on Telegram, a Channel; on Gab, a Group. Each community contains anywhere from a few to a few million users and is unrelated to "community" as used in network community detection.

To identify hateful communities for inclusion in our data, subject matter experts (SMEs) examined several thousands of candidate communities on each platform. SMEs checked each community's 20 most recent posts at some point during the year 2020. If at least 2 of these posts included hate speech, we included that community in our "hate universe" dataset.

The SMEs were generally in very high agreement. Here, by "hate speech", we refer to negative or hateful comments directed at a group or a class of persons on the basis of race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin [57,59].

We treated each community identified as a hate community as remaining so over time. Spot checks indicate that very few such communities stop posting hate speech in the long term, although their content is not always hateful.

Expression categories

We do not tailor the expression categories to our empirical context or our data set. Instead, we derive them from the existing theories of social identity formation. This ensures that the testing of said theories in our dataset is meaningful and does not yield a positive result via circular reasoning. For instance, a sense of shared grievance or injustice plays an important role in both SIMCA [30] and INN [2,40]. Mutual validation is a key mechanism that boosts the collective perception of shared grievance and translates it to action norms in both INN and the normative alignment model [31]. Shared efficacy belief plays important mediating roles in both the normative alignment model and SIMCA.

In contrast to the choice of categories, the specific regular expressions that represent them (see the Supplementary Text in S1 File) are rooted in our empirical context. We base these regular expressions on a prior experience with the "hate universe" database [41–43] (see also Materials and Methods), as well as a search through existing publications and other web-based sources.

A specific post can match more than one category of expression – for instance, both the grievance and the proposed action. In this case, we treat the contributions of such a post to the two types of expression as distinct. To put it another way, we assume that, for any given post, the probability of expression in one of the categories is independent of the probability of expression in another category. On the other hand,in designing our regular expression filters, we make an effort to ensure that the categories, as defined by these filters, are distinct – namely, that no two regular expressions from different categories match the same text (see the Supplementary Text in S1 File).

Our regular expressions could be missing some of the posts that fall in the categories that the posts intend to capture. Furthermore, according to a rather conservative, human-based verification procedure (see Supplementary Text in S1 File), about 30% of the posts that the regular expressions detect do not contain statements that fall in the corresponding category. Nevertheless, we posit that our time series generally track the expression categories that we want them to track, albeit with considerable noise. One qualitative piece of evidence in favor of this proposal is that the time series for the five categories appear to be punctuated by the election and the attack. In particular, maxima in the time series roughly correspond to these two events. Another piece of evidence is that the behaviors of the time series are distinctly different between the three time periods that are delimited by the events. The time series also appear to be different between categories, so that there is a possibility of a non-trivial effect of expression in one category on that in another, over time.

Time series

The magnitudes of post counts are substantially different between the different expression categories. For example, the posts with grievance expression are often several times more numerous than the posts that express belonging to the social identity. However, we wish to treat every category on an equal footing with every other. Specifically, we are interested in the way that proportional changes in one type of talk react to the proportional changes in another type of talk, not in the overall amounts of each type of talk. Accordingly, we scale each variable to have a mean of zero and a standard deviation of one across all the time points in a given period. The resulting time series, shown in S1 Fig, still differ qualitatively among the expression categories and among the three time periods. There, we observe that the propagated binomial error is smaller than the daily fluctuations in the scaled variables, even for *efficacy*, the expression category with the least numerous posts. In the remainder of this article, we work exclusively with the scaled versions of the five main variables.

Stationarity tests

We assess the stationarity of each variable during each time period via the Augmented Dickey-Fuller (ADF) test. This test is based on the following model of a single-valued time series:

$$\Delta y_t = \alpha + \beta t + \gamma t^2 + \rho y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i}, \tag{1}$$

where $\Delta y_t \equiv y_t - y_{t-1}$ and p is the lag order. The null hypothesis for the test is non-stationarity, $\rho = 0$. The alternative hypothesis is stationarity, $\rho < 0$.

We implement the ADF test via the `Python` function `adfuller` in package `statsmodels`, with the function's default option `autolag='AIC'`. We set the `regression` option to `'c'`, `'ct'`, or `'ctt'`; these option values correspond to degrees 0, 1 and 2 of the time polynomial in the above equation. Degree $d=0$ corresponds to $\beta = \gamma = 0$, $d=1$ corresponds to $\gamma = 0$, and $d=2$ – to a lack of any constraints on α, β or γ. Table 2 presents the *P*-values for the tests.

Whenever the ADF test is significant with $d=0$, we say that the series is stationary. When it is not significant with $d=0$ yet significant with either $d=1$ or $d=2$, we say that the series is trend-stationary. In this latter case, we "detrend" the variable, via the following procedure. We choose the smallest value of d where the series is trend-stationary. Subsequently, we build a regression of the focal variable on time with the appropriate degree – that is, a linear regression if $d=1$ or a quadratic regression if $d=2$. We fit the regression via the `Python` command `statsmodels.api.OLS`, with default options. We then subtract the regression equation from the variable, leaving only the residuals. These residuals constitute the detrended variable, which is stationary under $d=0$ for all our originally trend-stationary variables. In the one case where no $d \in \{0, 1, 2\}$ yields a stationary or a trend-stationary time series, we say that the variable is non-stationary. In this case, we difference the variable by substituting y_t with Δy_t; the latter is, once again, stationary under $d=0$. The resulting detrended and differenced variables subsequently enter into the vector auto-regression (VAR) model, alongside the variables that are stationary without detrending or differencing.

Table 2. Significance of the ADF test for stationarity in the 2020/2021 dataset.

Period 1					
Degree of trend	**Grievance**	**Efficacy**	**Validation**	**Identity**	**Action**
0	**0.0022**	**0.0301**	0.8444	0.6444	0.0954
1	0.0022	0.1292	**0.0000**	**0.0273**	0.1625
2	0.0004	0.0002	0.0000	0.0015	0.1927
Period 2					
Degree of trend	**Grievance**	**Efficacy**	**Validation**	**Identity**	**Action**
0	0.5552	**0.0004**	**0.0106**	**0.0034**	0.5242
1	**0.0132**	0.0000	0.0080	0.0517	**0.0041**
2	0.0782	0.0000	0.0091	0.0210	0.0068
Period 3					
Degree of trend	**Grievance**	**Efficacy**	**Validation**	**Identity**	**Action**
0	0.1405	0.5453	0.067	**0.0001**	**0.0000**
1	0.1920	0.7135	0.2589	0.0366	0.0007
2	**0.0034**	**0.0360**	**0.0000**	0.3445	0.0000

Each numeric entry is a P-value. For each column in each table, the P-value that is significant at the 5% level for the lowest-degree trend is in bold. We subtract a time regression with this degree from each variable's time series to obtain the variable's stationary portion, which then enters into the VAR analysis. For the variables where $P > 0.05$ for all trend degrees, we obtain the stationary variable by differencing.

https://doi.org/10.1371/journal.pcsy.0000107.t002

Collinearity measure

We assess collinearity by the means of the variance inflation factor (VIF). This factor is equal to $1/(1-R^2)$, where R^2 is the coefficient of determination for the linear regression of a focal variable on all the others. When VIF is above 5, this suggests that the focal variable is somewhat easy to predict via a linear regression from all the others. When VIF is above 10, the focal variable is very easy to predict from the others. High VIF indicates that the focal variable should be excluded in a dynamic model such as the vector auto-regression, where the goal is to establish temporal effects among independent components of the collective social identity. Specifically, when the lagged values of all five variables at some time are used to predict one of the variables at a later time in the context of a VAR model, collinearity makes it unclear which of the lagged variables contributes to the model's prediction. To put it another way, there can be simultaneous changes to several of the corresponding coefficients that do not change the prediction. This can contribute to an uncertainty in the estimation of the coefficients.

To calculate the Variance Inflation Factor (VIF), we use the `Python` function `variance_inflation_factor` in package `statsmodels.stats.outliers_influence`. The VIF values for the original variables are in the upper panel of Table 3. In the lower panel of this table, the variables that are not stationary have been transformed via detrending or differencing, as described above. In Period 3, the VIF for the *identity* variable is particularly high in both panels. Correspondingly, during this period, *identity* is strongly correlated with the other variables. Quantitatively speaking, for the original versions of the variables, $R^2 = 0.91$ in a multiple linear regression model with *identity* as the dependent variable and the remaining variables as independent. Thus, more than 90% of the variance in *identity* is explained by the combined variance of the other variables. The regression coefficients are all significant at the 5% significance level. Their values are 0.66 for *action*, 0.36 for *validation*, 0.16 for *efficacy* and −0.14 for *grievance*. It appears that, in Period 3, the expression of belonging to a common action-focused identity is correlated with the expression of all the associated components except, interestingly, *grievance* – the component that initiates the identity's development in the first place.

Accordingly, we remove *identity* from the analysis in Period 3. Table 3 indicates that upon this removal, the VIF values of all the variables that enter into the vector auto-regression are below 5.

Table 3. Collinearity in the 2020/2021 dataset.

Original variables					
	Grievance	**Efficacy**	**Validation**	**Identity**	**Identity**
Period 1	1.81	1.06	2.83	2.18	1.45
Period 2	3.62	1.93	3.27	1.46	1.39
Period 3	**5.80**	2.23	**8.25**	**11.09**	**8.06**
Stationary, detrended and differenced variables that enter into VAR					
	Grievance	**Efficacy**	**Validation**	**Identity**	**Action**
Period 1	1.22	1.11	1.54	1.39	1.03
Period 2	1.95	1.49	2.53	1.67	1.74
Period 3	1.66	1.19	1.59	**7.82**	**8.00**
Period 3 (no identity)	1.54	1.06	1.55	–	1.06

Each numeric entry is a value of the variance inflation factor (VIF). Top panel shows values for the original scaled variables. In the bottom panel, some of the variables are transformed to ensure stationarity, via either de-trending or differencing. Values greater than 5 are indicated in bold. Such a value corresponds to a variable that can be predicted from the remaining variables at any given time point. Upon the removal of *identity* from Period 3, VIF values for the remaining variables in the bottom panel are below 5.

https://doi.org/10.1371/journal.pcsy.0000107.t003

VAR and pathway analysis

Our statistical analysis of day-to-day influences among the different social identity components is based on the vector auto-regression (VAR) model. This model postulates that one can predict each variable's value at a given time step from its value – and the values of the remaining variables – at previous times. The model's formal definition is:

$$Y_t = \sum_{i=1}^{p} \Gamma_i Y_{t-i} + C + E_t, \tag{2}$$

where Y_t is a k-dimensional column vector (or, simply, k-vector) of variables Y at time t. In our case, each component of Y is the scaled proportion of posts that falls in one of the above-mentioned categories. Eq (2) also contains the regression intercepts, which are the elements of the constant k-vector C, as well as the model's residual errors at t, in the k-vector E_t. Index i is called lag and parameter p is the model's maximum lag (otherwise known as its lag order). The lag order determines how far back in time one has to look to predict Y in the present. The $k \times k$ matrix Γ_i contains the regression coefficients that determine the predictive interactions among the elements of Y at lag i. For instance, the diagonals of this matrix tell us how much the value of a given variable at time $t-i$ contributes to the value of that same variable at time t, for any t. The VAR model can be seen as a special case of a multiple, multivariate linear regression, in which the design matrix consists of the variables' lagged values. The model's traditional form assumes time-invariant, normal error distributions and zero covariance among errors across the different components of Y.

The VAR model seeks to predict the variables' current and future values from their values in the past. The model assumes that each variable is stationary – that is to say, that the variable's statistical properties, such as its mean and its standard deviation, do not change with time. If the assumption of stationarity does not hold for at least one of the variables, then a constant matrix Γ_i might not adequately describe the relationship among the variables at all times.

To fit the VAR model, we utilize the Stata command var.

To determine the optimal lag order of the VAR model for each period, we consider the final prediction error (FPE) and three information criteria versus lag. The information criteria are the Akaike information criterion (AIC), the Hannan-Quinn information criterion (HQIC), and Schwarz's Bayesian information criterion (SBIC). We compute FPE, AIC, HQIC and SBIC via the Stata command varsoc. S7 Fig plots these criteria versus lag order for each period. Here, HQIC and SBIC both clearly indicate that the optimal lag order is $p = 1$ for all three periods. AIC and FPE concur with HQIC and SBIC for Period 2; however, AIC and FPE are ambiguous for Period 3 and disagree with HQIC and SBIC for Period 1.

According to the "Remarks and Examples" section of [60], FPE is not strictly speaking, an information criterion. The same source states that the BIC and the HQIC have a theoretical advantage over the AIC and the FPE: choosing the lag order to minimize the BIC and the HQIC provides consistent estimates of the true lag order. In contrast, minimizing the AIC or the FPE will overestimate the true lag order with positive probability, even with an infinite sample size. In accordance with all of the above observations and theoretical considerations, as well as toward consistency in the statistical procedure across the periods, we choose $p = 1$, as recommended by HQIC and SBIC for all three periods.

The variables that enter into VAR are the stationary variants of all the original, scaled variables. For the variables that are stationary to begin with, the stationary versions are the same as the original scaled variables. For the variables that are trend-stationary, the stationary variants are the stationary portions that remain after detrending. Since *action* during Period 1 is neither stationary nor trend-stationary, its stationary variant is the original variable's differenced version. We also remove *identity* from the VAR analysis in Period 3, where this variable is highly collinear with the others.

In the pathway analysis of Period 2 in 2020/2021, one of the nodes, 4Chan's "Politically Incorrect" board (hereafter referred to as "4Chan1"), is a clear, extreme outlier in the distributions of the number of posts. Its posts number thousands to tens of thousands for each category, whereas the numbers for other communities are in tens to hundreds. We exclude 4Chan1 from the pathway analysis, since including it would reduce the effective number of observations to one.

The proportions and counts of posts that enter into the VAR and pathway analyses are associated with the above-mentioned false positive error, which we conservatively estimate to be on the order of ~30%. In the context of regression analyses, this error on individual points – either time points in the case of VAR or community points in the case of pathway analysis – translates to appreciably lower errors on the regression coefficients when the number of points is large. For example, in the case of Period 2 VAR analysis, each coefficient is estimated from $n \approx 60$ points and the variance of the predictor, i.e., lagged, variable is $\mathrm{Var}(x) \approx 1$. If the standard deviation due to the false positive error is $\sigma \approx 0.3^2$, then the error in the regression slope is approximately $\sigma / \sqrt{n \times \mathrm{Var}(x)} \approx 0.04$. This error is no more than half the magnitudes of the smallest coefficients that we estimate as significant; these latter magnitudes are on the order of 0.1. For the majority of our coefficient estimates, the relative error due to the false positives in the regular expressions should be appreciably lower than this conservative estimate of about half the coefficient's magnitude. Accordingly, if the statistical properties of the false positives are such that they do not introduce appreciable bias to our qualitative conclusions, we estimate that the potential effect of the false positives is, at most, to invalidate the significance of a few of the lowest-magnitude regression coefficients. We further point out that there is an inherent trade-off between the transparency and the perceived accuracy of natural language processing tools that propose to operationalize abstract psychological concepts in new data. Accordingly, our relatively simple regular expression procedure offers a methodology that is complementary to advanced tools such as large language models.

2024 election

The two grayed out intervals in Fig 5 are marked by instability in 4Chan, the social media platform that accounts for a large percentage of the total dataset. Due to this fact, the data during these intervals is incomplete. Furthermore, the data before the first grayed out interval was collected using a different sampling method than the data after this interval. Accordingly, we do not include the first grayed out interval in our analysis.

For the Period 1b VAR model, the data-to-parameter ratio is relatively low – only about 7.3 when the lag order p is 1; this ratio is even lower for higher lag orders. Accordingly, there is considerable uncertainty with regard to the optimal lag order, as evident in the middle panel of S8 Fig. Additionally, although three of the four information measures in this figure favor high lag orders by decreasing up to $p = 5$, Stata's fits for $p > 5$ are not reliable. Thus, we conclude that the Period 1b data is insufficient to confidently fit the auto-regressive model. An implicit assumption of our analysis of the 2020 election cycle is that the rules of the system's behavior are time-uniform before the election day. If we apply the same assumption to the 2024 cycle, then the absence of the data in the first grayed out interval of Fig 5 and the data in Period 1b only detracts from the statistical power of our conclusions regarding the pre-election behavior as inferred from Period 1a, without altering the qualitative nature of these conclusions.

Supporting information

S1 File. Supplementary text. The supplementary text includes information on the following topics: neutrality of the research design, regular expressions, verification of categories, impulse response functions, pathway analysis, 2024 election analysis, comparison of themes between 2020 and 2024.
(PDF)

S1 Fig. Scaled proportions of posts in the 2020/2021 dataset, versus time.
(PDF)

S2 Fig. Impulse response functions (IRF) for the vector auto-regression (VAR) in Period 1 (2020).
(PDF)

S3 Fig. Impulse response functions (IRF) for the vector auto-regression (VAR) in Period 2 (2020/2021).
(PDF)

S4 Fig. Impulse response functions (IRF) for the vector auto-regression (VAR) in Period 3 (2021).
(PDF)

S5 Fig. Post counts that enter the pathway analysis for Period 2 (2020/2021) communities.
(PDF)

S6 Fig. Scaled proportions of posts in the 2024/2025 dataset, versus time.
(PDF)

S7 Fig. Comparison of VAR models with different lag orders in the analysis of the 2020/2021 data.
(PDF)

S8 Fig. Comparison of VAR models with different lag orders in the analysis of the 2024/2025 data.
(PDF)

S1 Table. Interval boundaries in the pathway analysis of Period 2 (2020/2021).
(PDF)

S2 Table. Significance of the ADF test for stationarity in the 2024/2025 dataset.
(PDF)

S3 Table. Collinearity in the 2024/2025 dataset.
(PDF)

S4 Table. Top themes identified by ChatGPT in the 2020 and 2024 posts.
(PDF)

Author contributions

Conceptualization: Mikhail Lipatov, Lucia Illari, Akshay Verma, Neil F. Johnson, Sergey Gavrilets.

Data curation: Mikhail Lipatov, Lucia Illari, Richard Sear, Akshay Verma, Neil F. Johnson.

Formal analysis: Mikhail Lipatov, Akshay Verma, Neil F. Johnson, Sergey Gavrilets.

Funding acquisition: Neil F. Johnson, Sergey Gavrilets.

Investigation: Mikhail Lipatov, Lucia Illari, Richard Sear, Akshay Verma, Neil F. Johnson, Sergey Gavrilets.

Methodology: Mikhail Lipatov, Lucia Illari, Richard Sear, Akshay Verma, Neil F. Johnson, Sergey Gavrilets.

Project administration: Mikhail Lipatov, Neil F. Johnson, Sergey Gavrilets.

Resources: Mikhail Lipatov, Richard Sear, Neil F. Johnson, Sergey Gavrilets.

Software: Mikhail Lipatov, Lucia Illari, Richard Sear.

Supervision: Neil F. Johnson, Sergey Gavrilets.

Validation: Mikhail Lipatov, Lucia Illari, Richard Sear, Neil F. Johnson, Sergey Gavrilets.

Visualization: Mikhail Lipatov, Richard Sear, Akshay Verma.

Writing – original draft: Mikhail Lipatov, Sergey Gavrilets.

Writing – review & editing: Mikhail Lipatov, Richard Sear, Neil F. Johnson, Sergey Gavrilets.

References

1. McGarty C, Thomas EF, Lala G, Smith LG, Bliuc AM. New technologies, new identities, and the growth of mass opposition in the Arab spring. Polit Psychol. 2014;35(6):725–40.

2. Smith LGE, Gavin J, Sharp E. Social identity formation during the emergence of the occupy movement. Euro J Social Psych. 2015;45(7):818–32. https://doi.org/10.1002/ejsp.2150
3. Smith LG, McGarty C, Thomas EF. After Aylan Kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time. Psychol Sci. 2018;29(4):623–34.
4. Thomas EF, Smith LGE, McGarty C, Reese G, Kende A, Bliuc A, et al. When and how social movements mobilize action within and across nations to promote solidarity with refugees. Eur J Soc Psychol. 2018;49(2):213–29. https://doi.org/10.1002/ejsp.2380
5. Blum I, Uldam J. Faking, optimising and conceding to power: Social movement understandings of social media power. New Media Soc. 2024;27(11):6233–51. https://doi.org/10.1177/14614448241266769
6. Lee FL, Fong IW. The construction and mobilization of political consumerism through digital media in a networked social movement. New Media Soc. 2021;25(12):3573–92. https://doi.org/10.1177/14614448211050885
7. Steinert-Threlkeld ZC, Mocanu D, Vespignani A, Fowler J. Online social networks and offline protest. EPJ Data Sci. 2015;4(1). https://doi.org/10.1140/epjds/s13688-015-0056-y
8. Pinckney J, Trilling C. Breaking down pillars of support for democratic backsliding. Mobilization. 2024.
9. Stammen L, Meissner M. Social movements' transformative climate change communication: extinction rebellion's artivism. Soc Mov Stud. 2022;23(1):19–38. https://doi.org/10.1080/14742837.2022.2122949
10. Wimmer A, Torrats-Espinosa G. COVID, compassion and altruistic mobilization: Explaining non-Black participation in the Black Lives Matter movement of 2020. Mobilization. 2025.
11. Lobbedez E, Buchter L. The strength of pushback collective identity in a fragmented mass movement*. Mobilization: Int Quarterly. 2023;28(1):61–88. https://doi.org/10.17813/1086-671x-28-1-61
12. Tullock G. The paradox of revolution. Public Choice. 1971;11(1):89–99. https://doi.org/10.1007/bf01726214
13. Granovetter M. Threshold models of collective behavior. Am J Sociol. 1978;83(6):1420–43. https://doi.org/10.1086/226707
14. Kuran T. Sparks and prairie fires: a theory of unanticipated political revolution. Public Choice. 1989;61(1):41–74. https://doi.org/10.1007/bf00116762
15. Gavrilets S. Collective action problem in heterogeneous groups. Philos Trans R Soc Lond B Biol Sci. 2015;370(1683):20150016. https://doi.org/10.1098/rstb.2015.0016 PMID: 26503689
16. Gavrilets S, Tverskoi D, Wang N, Wang X, Ozaita J, Zhang B, et al. Co-evolution of behaviour and beliefs in social dilemmas: estimating material, social, cognitive and cultural determinants. Evol Hum Sci. 2024;6:e50. https://doi.org/10.1017/ehs.2024.38 PMID: 39703942
17. Gavrilets S. Social influence and the logic of collective action. Princeton, NJ: Princeton University Press; 2026.
18. Watts DJ. A simple model of global cascades on random networks. Proc Natl Acad Sci U S A. 2002;99(9):5766–71. https://doi.org/10.1073/pnas.082090499 PMID: 16578874
19. Centola D. The spread of behavior in an online social network experiment. Science. 2010;329(5996):1194–7. https://doi.org/10.1126/science.1185231 PMID: 20813952
20. Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, et al. A 61-million-person experiment in social influence and political mobilization. Nature. 2012;489(7415):295–8. https://doi.org/10.1038/nature11421 PMID: 22972300
21. Korkmaz G, Kuhlman CJ, Goldstein J, Vega-Redondo F. A computational study of homophily and diffusion of common knowledge on social networks based on a model of Facebook. Soc Netw Anal Min. 2019;10(1). https://doi.org/10.1007/s13278-019-0615-5
22. Constantino SM, Sparkman G, Kraft-Todd GT, Bicchieri C, Centola D, Shell-Duncan B, et al. Scaling up change: a critical review and practical guide to harnessing social norms for climate action. Psychol Sci Public Interest. 2022;23(2):50–97. https://doi.org/10.1177/15291006221105279 PMID: 36227765
23. Wan A, Riedl C, Lazer D. Diffusion of complex contagions is shaped by a trade-off between reach and reinforcement. Proc Natl Acad Sci U S A. 2025;122(28):e2422892122. https://doi.org/10.1073/pnas.2422892122 PMID: 40638089
24. Tajfel H, Turner J, Austin WG, Worchel S. An integrative theory of intergroup conflict. Intergroup relations: Essential readings. 2001. pp. 94–109.
25. Turner JC, Hogg MA, Oakes PJ, Reicher SD, Wetherell MS. Rediscovering the Social Group: A Self-Categorization Theory. Oxford: Basil Blackwell; 1987.
26. Festinger L. A theory of social comparison processes. Hum Relat. 1954;7(2):117–40. https://doi.org/10.1177/001872675400700202
27. Haslam SA, Turner JC, Oakes PJ, McGarty C, Reynolds KJ. The group as a basis for emergent stereotype consensus. Eur Rev Soc Psychol. 1997;8(1):203–39. https://doi.org/10.1080/14792779643000128
28. Postmes T, Spears R, Lee AT, Novak RJ. Individuality and social influence in groups: inductive and deductive routes to group identity. J Pers Soc Psychol. 2005;89(5):747–63. https://doi.org/10.1037/0022-3514.89.5.747 PMID: 16351366
29. Postmes T, Haslam SA, Swaab RI. Social influence in small groups: An interactive model of social identity formation. Eur Rev Soc Psychol. 2005;16(1):1–42. https://doi.org/10.1080/10463280440000062
30. van Zomeren M, Postmes T, Spears R. Toward an integrative social identity model of collective action: a quantitative research synthesis of three socio-psychological perspectives. Psychol Bull. 2008;134(4):504–35. https://doi.org/10.1037/0033-2909.134.4.504 PMID: 18605818

31. Thomas EF, McGarty C, Mavor KI. Aligning identities, emotions, and beliefs to create commitment to sustainable social and political action. Pers Soc Psychol Rev. 2009;13(3):194–218.
32. Milesi P, Alberici AI. Pluralistic morality and collective action: the role of moral foundations. Group Process Intergroup Relat. 2016;21(2):235–56. https://doi.org/10.1177/1368430216675707
33. Thomas EF, McGarty C, Louis W. Social interaction and psychological pathways to political engagement and extremism. Euro J Social Psych. 2013;44(1):15–22. https://doi.org/10.1002/ejsp.1988
34. Smith LGE, Piwek L, Hinds J, Brown O, Chen C, Joinson A. Digital traces of offline mobilization. J Pers Soc Psychol. 2023;125(3):496–518. https://doi.org/10.1037/pspa0000338 PMID: 36780273
35. Wischerath D, Godwin E, Bocheva D, Brown O, Roscoe JF, Davidson BI. Spreading the Word: Exploring a Network of Mobilizing Messages in a Telegram Conspiracy Group. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 2024. pp. 1–8.
36. Brown O, Smith LG, Davidson BI, Racek D, Joinson A. Online signals of extremist mobilization. Pers Soc Psychol Bull. 2024. https://doi.org/10.1177/01461672241266866
37. Brown O, Smith LG, Davidson BI, Racek D, Joinson A. Online risk signals of offline terrorist offending. OSF; 2023.
38. Haslam SA, Reicher SD. When prisoners take over the prison: a social psychology of resistance. Pers Soc Psychol Rev. 2012;16(2):154–79. https://doi.org/10.1177/1088868311419864 PMID: 21885855
39. Tajfel H, Billig MG, Bundy RP, Flament C. Social categorization and intergroup behaviour. Eur J Soc Psychol. 1971;1(2):149–78. https://doi.org/10.1002/ejsp.2420010202
40. Smith LGE, Thomas EF, McGarty C. "We Must Be the Change We Want to See in the World": Integrating norms and identities through social interaction. Polit Psychol. 2014;36(5):543–57. https://doi.org/10.1111/pops.12180
41. Johnson NF, Leahy R, Restrepo NJ, Velasquez N, Zheng M, Manrique P, et al. Hidden resilience and adaptive dynamics of the global online hate ecology. Nature. 2019;573(7773):261–5. https://doi.org/10.1038/s41586-019-1494-7 PMID: 31435010
42. Zheng M, Sear RF, Illari L, Restrepo NJ, Johnson NF. Adaptive link dynamics drive online hate networks and their mainstream influence. npj Complex. 2024;1(1). https://doi.org/10.1038/s44260-024-00002-2
43. Verma A, Sear R, Johnson N. How U.S. Presidential elections strengthen global hate networks. Npj Complex. 2024;1(1):18. https://doi.org/10.1038/s44260-024-00018-8 PMID: 39498374
44. Tajfel H. Human groups and social categories: Studies in social psychology. Cambridge University Press; 1981.
45. Runciman WG. Relative Deprivation and Social Justice: A Study of Attitudes to Social Inequality in Twentieth-century England. Reports of the Institute of Community Studies. University of California Press; 1966. https://books.google.com/books?id=1OfEAAAAIAAJ
46. Smith LGE, Postmes T. Intra-group interaction and the development of norms which promote inter-group hostility. Eur J Soc Psychol. 2009;39(1):130–44. https://doi.org/10.1002/ejsp.464
47. Smith LGE, Postmes T. The power of talk: developing discriminatory group norms through discussion. Br J Soc Psychol. 2011;50(Pt 2):193–215. https://doi.org/10.1348/014466610X504805 PMID: 21545454
48. Johnson N, Carran S, Botner J, Fontaine K, Laxague N, Nuetzel P, et al. Pattern in escalations in insurgent and terrorist activity. Science. 2011;333(6038):81–4. https://doi.org/10.1126/science.1205068 PMID: 21719677
49. Johnson NF, Medina P, Zhao G, Messinger DS, Horgan J, Gill P, et al. Simple mathematical law benchmarks human confrontations. Sci Rep. 2013;3:3463. https://doi.org/10.1038/srep03463 PMID: 24322528
50. Bohorquez JC, Gourley S, Dixon AR, Spagat M, Johnson NF. Common ecology quantifies human insurgency. Nature. 2009;462(7275):911–4. https://doi.org/10.1038/nature08631 PMID: 20016600
51. Kann C, Hashash S, Steinert-Threlkeld Z, Alvarez RM. Collective identity in collective action: evidence from the 2020 summer BLM protests. Front Polit Sci. 2023;5. https://doi.org/10.3389/fpos.2023.1185633
52. Hess D, Martin B. Repression, backfire, and the theory of transformative events. Mobiliz: Int Quarterly. 2006;11(2):249–67.
53. Fu D, Göbel C. Exposing state repression: digital discursive contention by Chinese protestors. Stud Comp Int Dev. 2025;60(3):655–89. https://doi.org/10.1007/s12116-024-09428-0 PMID: 41140540
54. Falkenberg M, Zollo F, Quattrociocchi W, Pfeffer J, Baronchelli A. Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. Nat Commun. 2024;15(1):9560. https://doi.org/10.1038/s41467-024-53868-0 PMID: 39543121
55. Manekin D, Mitts T, Zeira Y. The politics of allyship: Multiethnic coalitions and mass attitudes toward protest. Proc Natl Acad Sci U S A. 2024;121(19):e2314653121. https://doi.org/10.1073/pnas.2314653121 PMID: 38696470
56. Shuman E, Hasan-Aslih S, van Zomeren M, Saguy T, Halperin E. Protest movements involving limited violence can sometimes be effective: evidence from the 2020 BlackLivesMatter protests. Proc Natl Acad Sci U S A. 2022;119(14):e2118990119. https://doi.org/10.1073/pnas.2118990119 PMID: 35344420
57. Lupu Y, Sear R, Velásquez N, Leahy R, Restrepo NJ, Goldberg B, et al. Offline events and online hate. PLoS One. 2023;18(1):e0278511. https://doi.org/10.1371/journal.pone.0278511 PMID: 36696388

58. Velásquez N, Leahy R, Restrepo NJ, Lupu Y, Sear R, Gabriel N, et al. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. Sci Rep. 2021;11(1):11549. https://doi.org/10.1038/s41598-021-89467-y PMID: 34131158

59. American Library Association. Hate Speech and Hate Crime. 2025. [cited 2025 Sep 17]. Available from: https://www.ala.org/advocacy/intfreedom/hate

60. StataCorp. VAR and VEC estimation and postestimation: varsoc – Lag-order selection statistics. Stata Time-Series Reference Manual Release 17. College Station, TX. 2021. Available from: https://www.stata.com/manuals/tsvarsoc.pdf