PNAS Nexus, 2024, 3, pgae534
https://doi.org/10.1093/pnasnexus/pgae534
Advance access publication 22 November 2024
Research Report

Sanctioning political speech on social media is driven by partisan norms and identity signaling

Chloe Ahn[a,b], Yphtach Lelkes[a] and Matthew Levendusky[b,c,*]

[a]Annenberg School for Communication, University of Pennsylvania, 3620 Walnut St, Philadelphia, PA 19104, USA
[b]Department of Political Science, University of Pennsylvania, 133 S. 36th St, Philadelphia, PA 19104, USA
[c]Annenberg Public Policy Center, University of Pennsylvania, 202 S. 36th St, Philadelphia, PA 19104, USA
*To whom correspondence should be addressed: Email: mleven@sas.upenn.edu
Edited By Erik Kimbrough

Abstract

Social media is marked by online firestorms where people pile-on and shame those who say things perceived to be offensive, especially about politically relevant topics. What explains why individuals engage in this sort of sanctioning behavior? We show that two key factors help to explain this pattern. First, on these topics, both offensive speech and subsequent sanctioning are seen as informative about partisanship: people assume that those who say offensive things are out-partisans, and those who criticize them are co-partisans. Second, individuals perceive that such sanctioning is an injunctive norm and believe that their fellow co-partisans approve of it—sanctioning someone allows them to signal their partisanship by adhering to that norm. Using three original experiments, we show strong support for this argument. Sanctioning this type of offensive speech is as informative about perceived partisanship as explicit partisan electioneering. Further, people perceive that a wide variety of sanctioning behaviors are (partisan) group norms. We also show that while people are reluctant to be the first to criticize someone online, they are quite willing to pile-on to others' criticisms, which helps to explain why this behavior spreads so rapidly in online firestorms. Our results have implications for online social dynamics, as well as partisanship and partisan animosity more broadly.

Keywords: cancellation, social media, group norms, identity signaling, political speech

Significance Statement

Why do individuals sanction politically offensive speech online? We argue that it allows them to signal their partisanship and to adhere to perceived partisan group norms. Using several original experiments, we show that people view sanctioning offensive speech as a signal of partisanship similar to electioneering. While many people are hesitant to be the first person to sanction someone's offensive speech, they will readily pile-on to someone else's criticism. These dynamics help to explain the prevalence of this type of behavior online and why it can spread so rapidly there.

Introduction

In 2022, the *New York Times* declared that Americans were "losing hold of a fundamental right as citizens of a free country: the right to speak their minds and voice their opinions" due to the proliferation of social and economic sanctioning of unpopular speech—colloquially referred to call-out or cancel culture (1).[a] This occurs when someone posts something offensive online, which then triggers responses ranging from critical comments, to boycotts, and even doxing (2). Such behavior spreads rapidly once it starts, leading to online firestorms.

The social sanctioning of political speech online occurs because it allows individuals to demonstrate their political loyalties and conform to their group's norms (3). Such social sanctioning is typically a response to highly moralized political issues such as racism, gender identity, and so forth (2). But, it is also performative (4). People sanction others to signal their partisan identity and adhere to group norms (5, 6). Because the parties have taken distinctive stances on these issues over time, and these types of morality-infused issues defy compromise (7), they are excellent tools for "moral grandstanding": position-taking used to signal one's identity and improve in-group (partisan) esteem (8). So while the general idea of identity signaling is well-established in many areas, from psychology (3, 4), to economics (9, 10), to political science (11), our contribution is to demonstrate this same mechanism applies to this sort of online policing.

Throughout our argument, we focus only on political speech: that is, topics that are plausibly connected to politics, particularly in the United States. While this is an important scope condition, because politics has come to infuse even many apolitical behaviors (12), and so many actions and statements are perceived to

Competing Interest: The authors declare no competing interests.
Received: September 6, 2024. **Accepted:** November 8, 2024

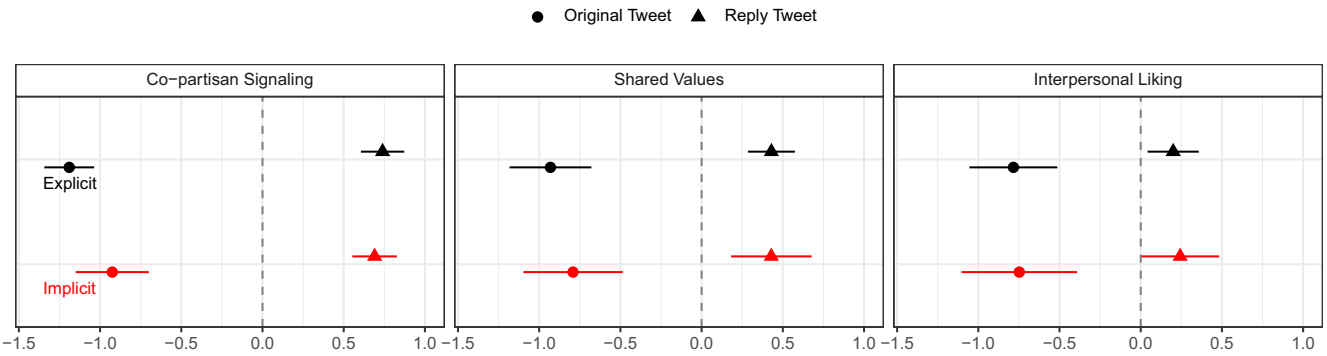

Fig. 1. Respondents' assessment of (1) those who write either implicit (red circles) or explicit political (black circles) tweets, relative to apolitical tweets, and (2) those who criticize such implicit (red triangles) or explicit political (black triangles) tweets, relative to those who respond to apolitical tweets. All coefficients, with their 95% CI, are derived from fixed-effects OLS models, with tweet issue types as fixed effects and standard errors clustered by respondent.

be political (13), this is not an overly restrictive limitation. Indeed, most "real-world" cases of social sanctioning online focus on exactly these sorts of issues (2).

We also draw a distinction between two types of political speech: explicitly partisan speech, which directly addresses parties or candidates (for example, posts praising or attacking a particular candidate), and implicit partisan speech, which addresses topics where the parties take distinct positions but does not include an overt partisan cue (for example, discussions of transgender identity or racism). As we demonstrate, both powerfully shape how people react online. Indeed, sanctioning implicit speech is functionally as powerful of a signal of partisan identity as sanctioning explicit speech, underscoring the extent to which sanctioning behaviors have become enmeshed in partisanship.

To be clear, our argument is not that moralistic punishment is novel; that has existed for eons. Our argument instead is that social media provides new avenues for this behavior, allowing sanctioning to be relatively low-cost while having a wider reach and garnering peer approval (e.g. (14)). Moralized content is highly engaging and emotionally arousing, and gets algorithmically amplified (15, 16). Because social media is *social*, when an individual sanctions someone online, they receive praise from their peers (e.g. likes, shares, retweets), which signals that such sanctioning is normatively desirable (15, 17). Social media amplifies age-old dynamics.

But if this behavior pays social dividends, why has past research found that few people are willing to engage in it (2)? Group norms, even related to partisan speech, are not always clear and being the first person to sanction someone entails some risk: the individual who initiates the sanction might face personal attacks or ostracism if they misjudged the group norms. However, once others take that step, the risk diminishes while the reward (in terms of in-group signaling) remains significant. Although individuals may be hesitant to act initially, once someone else does, others quickly follow, facilitating the spread of this behavior online.

These processes have a number of testable empirical implications. First, we expect that sanctioning others' implicitly or explicitly political speech is a signal of partisan identity—people infer partisanship from sanctioning behavior online. Second, we expect that sanctioning will be perceived as an injunctive group norm: people think their fellow partisans will approve of efforts to sanction others' political speech online. Third, we expect that while most individuals are unwilling to be the first person to sanction someone online, they will be likely to pile on to someone else's efforts.

We test these hypotheses with three preregistered experiments that asked respondents to assess identity signaling and examine how group norms may influence support for sanctioning behavior. The results of our studies suggest that sanctioning behavior is a strategic display of political allegiance and adherence to group norms.

Results

Our first experiment asks if sanctioning sends a partisan signal. Respondents were asked to evaluate a statement and a reply, presented as tweets. Subjects were randomly assigned to one of three conditions: offensive implicitly partisan tweets and critical replies (real-world examples of tweets that are ideologically incongruent or morally unacceptable from the respondent's point of view, but do not mention parties), explicitly partisan tweets from the other party and replies (advocating for an out-party candidate, with replies that then criticize that candidate), and apolitical tweets and replies (about movies, the weather, and so forth). For each pair of statements, respondents were asked to assess the tweet's author along a variety of different dimensions. All data for this article can be found in Ref. (18).

Implicitly, partisan tweets signal partisanship and affect interpersonal attitudes nearly as strongly as explicitly partisan tweets. Seeing either an offensive implicitly partisan tweet, or an explicitly partisan tweet from the other party, substantially reduced perceptions that the author shared the respondent's partisanship (Fig. 1, circles in the left panel). For example, when a Democrat reads a tweet critical of transgender people, they think that the author is almost as likely to be a Republican as when they read about someone campaigning for Republican candidates, relative to when they read an anodyne apolitical tweet. For example, on our 1–5 scale of perceived co-partisanship (where higher values indicate a greater likelihood of shared partisanship), apolitical tweet scored (on average) at 3.040, implicitly political tweets at 1.907, and explicitly political tweets at 1.578. It is unsurprising that explicit partisan messages cue partisanship, it is much more noteworthy that implicit ones are nearly as effective at doing so.

Both implicitly and explicitly partisan messages also influence a respondent's perceptions of whether the author shares their values and has a lot in common with them (the middle panel) as well as their overall feelings towards them (right panel).

Crucially, respondents also draw a parallel set of inferences about those who sanction others online. If that same Democrat sees someone critiquing an implicitly Republican tweet, they think that critic is a Democrat (red triangles in the left panel). Indeed, they are just as likely to think that person is a Democrat as when they read about that person criticizing Republican candidates (black triangles in the left panel).

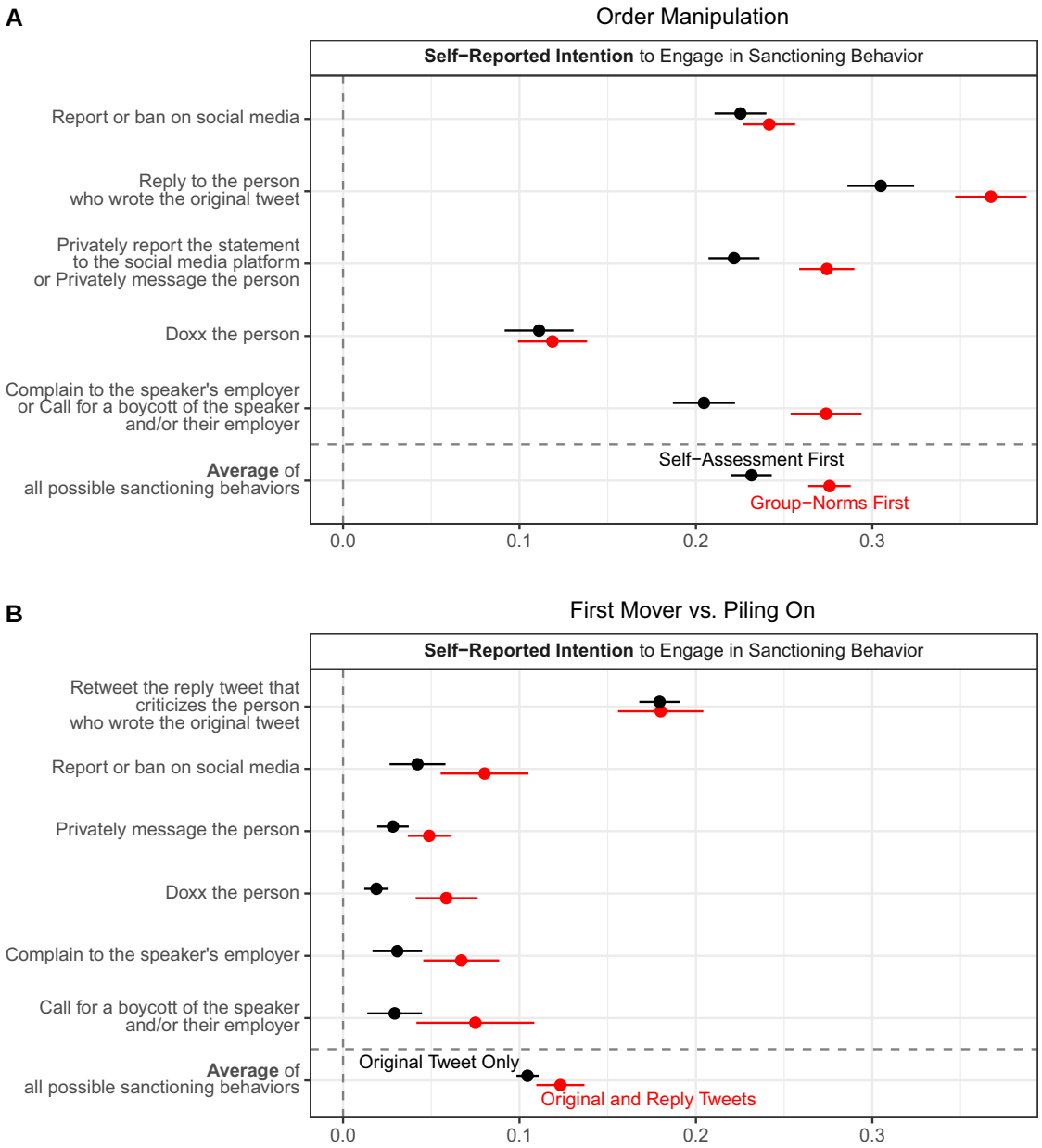

Fig. 2. A) The likelihood of engaging in sanctioning behaviors based on whether group norms are asked before (red circles) or after (black circles) individual actions are shown. B) The likelihood of sanctioning, based on whether the respondent needs to be the first mover (black circles) or can "pile on" to someone else's reply (red circles) is shown. In both panels, 95% CI are shown as lines. All models include tweet issue types as fixed effects, and the standard errors are clustered by respondent ID. Logistic regression models are used to analyze binary outcomes (all individual sanctioning behaviors), and OLS models are used when the outcome variable is continuous (average of all possible sanctioning behaviors).

To test whether group norms causally influence sanctioning behavior, we randomized the order in which respondents were asked (i) how they themselves would respond to offensive implicitly partisan statements online (which we assume they would perceive to be from the other party, given study 1 above) and (ii) whether their co-partisans would approve of them if they sanctioned these statements. This manipulation primed the in-group's expectations for those who answered co-partisan assessments first, and hence tested the effect of these norms on behavior (19).

When self-assessments were asked first, respondents were more reluctant to partake in any form of sanctioning behavior (black circles in Fig. 2A). However, when group norms were asked first, the likelihood of sanctioning increased (red circles in Fig. 2A). Respondents think sanctioning is a partisan norm, and hence engaging in it allows them to signal their group allegiance.

But Fig. 2A also highlights that most people are not especially willing to sanction others, consistent with earlier studies (2). If this is a norm, why is it not a more common behavior?

We argue that people are reluctant to be the first person to sanction someone but will pile on when others do it first. Acting first exposes people to the risk of push-back, but that is lessened when one is part of a group. Study 3 tests this logic by asking respondents how they would respond to an offensive implicitly partisan tweet. We randomly assigned them to see either the implicitly partisan tweet by itself, forcing them to be the first mover in critiquing it, or to see it with critical replies, allowing them to pile-on to someone else's social sanctioning efforts.

Figure 2B confirms this logic. When shown an implicitly partisan tweet with no replies (shown with black circles in Fig. 2B), forcing respondents to be the first mover, they were less likely to

sanction others than when shown that same tweet with critical replies, allowing them to pile-on to someone else's criticism (red circles in Fig. 2B). This was true across multiple forms of sanctioning behavior. Even if individuals will not be the first to criticize someone, they will pile-on once others have acted.

Discussion

Across three experiments, we demonstrate three key findings (i) social sanctioning of political speech is an informative partisan signal: those who write offensive implicitly or explicitly political tweets are perceived as out-partisans, while those who criticize them (or pile on to that criticism) are perceived as co-partisans, (ii) partisans believe that such sanctioning is a group norm, and (iiii) most people are reluctant to be the first person to sanction someone else, but they will pile-on once others have acted.

Together, these findings help to explain why we see social sanctioning online and endless discussions of "cancel culture." Even if few people want to sanction others (2), perceived group norms and opportunities to signal one's partisan bona fides beget piling on. Social media algorithms subsequently amplify this content, which only strengthens the norm and reinforces the cycle. Even if many people do not approve of this behavior (2, 20), it persists precisely because of these feedback dynamics (21).

The observation that sanctioning others' political speech online strongly signals partisanship underscores how this behavior can also contribute to partisan animosity more broadly. People assume that those who say offensive things are out-partisans, and those who critique them are co-partisans, so this behavior is seen through partisan-colored lenses. Even those who do not use social media themselves can fall prey to this logic because these behaviors are covered so extensively in the mass media (2), especially given two-step information flows (22). This is yet another way the nature of these platforms moralizes and polarizes our politics.

As with all experiments, our results suggest important directions for future studies. Here, while we focused on *politically* offensive speech, considering other cases is an important extension—do the same dynamics apply to other types of offensive speech or behavior? While this political process is typical of much social sanctioning online, it is not the totality of such behavior, and other cases may have different dynamics.

Further, while we focused on salient issues here (given how we selected our cases, see the materials section below), more obscure issues may have a different logic. Most posts on social media do not go viral, and individuals may calculate that the reward for sanctioning will be low if the issue lacks relevance or common knowledge among co-partisans, something future studies should investigate.

Finally, future studies can more directly test the effects of group norms by assessing how respondents judge other group members' sanctioning actions. Additionally, future research could also investigate whether participants believe they will be punished for not sanctioning or rewarded for doing so, providing a clearer understanding of injunctive group norms and their influence on sanctioning behavior. That said, even with these limitations, our findings still demonstrate important dynamics of social sanctioning online.

Materials and methods

All three experiments were preregistered with OSF prior to fielding; preregistrations, treatment details, and full questionnaires for all studies are available at: https://osf.io/v23x5/. Experiments for this study were reviewed by the University of Pennsylvania Institutional Review Board and deemed to be exempt.

All experiments were conducted on nonprobability samples provided through Bovitz Forthright. Study 1 was fielded between 2023 December 8th to 19th, $N = 1{,}477$; Study 2 was fielded between 2023 April 4th to 15th, $N = 2{,}470$; Study 3 was fielded between 2023 September 27th to October 10th, $N = 1{,}508$. We restricted our analysis to Democratic and Republican respondents (including leaning independents) given the nature of our hypotheses. The few pure independent respondents recruited were removed from our analysis. Responses were analyzed with OLS regressions, with clustering at the statement and respondent level; details on the methods are provided in the Supplementary materials.

In study 1, each subject was shown four different statements in the form of screenshots of tweets: three offensive or electioneering tweets and one apolitical tweet for measuring baseline behavioral intention. For the offensive tweets, we used the same set of tweet treatments as in study 2 (see below) but selected five tweet treatments that demonstrated the highest probabilities of eliciting sanctioning behavior from respondents. For the electioneering tweets, we sampled five tweets from the top 100 liked tweets from the 2020 presidential election tweet dataset; for apolitical tweets, we sampled five tweets from the top retweeted tweets in 2022 and manually added mundane apolitical tweets related to weather or pop culture.

In study 2, each subject was shown two different statements that had led to someone being sanctioned in the real world. Public comments involving offensive statements were collected from *USA Today* news articles, the nation's largest-circulation newspaper, obtained from the Nexis Uni database. We used a set of keywords related to social sanctioning and controversial statements (details provided in the preregistration). We sampled 12 statements where left-wing individuals sanctioned someone who said something more right-wing, and 10 with the opposite ideological valence (i.e. someone from the right-wing sanctioning a left-wing comment). For each tweet treatment, individuals were block-randomized based on the level of perceived extremity and offensiveness of that particular statement to match one of the sanctioning statements, which had been pretested in a pilot study.

We used ChatGPT to edit the statements to remove the speaker's identities and make the statements more general. Subjects were asked how they would respond to these statements if they saw them on social media, using the list of sanctioning ("canceling") behaviors from (2). They were also asked how their co-partisans would evaluate them if they engaged in each activity, which serves as a measure of injunctive partisan norms. To motivate respondents to accurately report these norms, we incentivized them with $1 if they accurately predicted their fellow co-partisans' approval of engaging in sanctioning behaviors (i.e. correct predictions of group norms). The experimental manipulation involved the order in which these items were asked: one-half of the participants were asked the self-reports first, and the other half were asked about co-partisan norms first (19). This study also included a partisan identity priming, though this did not affect responses; see the supporting information for details.

Study 3 asks whether people will pile one to someone else's efforts to sanction others. Following study 2, to find offensive statements that had been sanctioned, we used newspaper coverage of this topic in *USA Today*. Subjects were randomly assigned to one of two conditions: they were either simply shown the offensive statement, or they were shown the statement along with a reply (the reply criticizing the actual offensive statement online). Consistent with the previous study, all the sanctioned speakers

were high-profile individuals (2). However, because those criticizing them were not, we randomized their names to prevent their race or gender from affecting the outcome, using the list of validated names as described in the preanalysis plan. We were also concerned that the popularity (likes/shares) of the reply might influence respondents' behavior, so we randomized the popularity of each original statement and reply (see preregistration). However, as we show in the supporting information, this manipulation did not result in a meaningful difference in treatment effects, except when comparing highly popular reply tweets to those without any popularity information.

Note

[a] We use "social sanctioning" to refer to this behavior, rather than "canceling," as the latter has a highly contested and unclear meaning.

Acknowledgments

The authors thank Elias Dinas, Jamie Druckman, Justin Grimmer, Neil Malhotra, Antoine Marie, Jesper Rasmussen, the editor, the anonymous referees, and seminar participants at the Polarization Research Lab Conference, the European University Institute, and the University of Pennsylvania for helpful comments.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

There are no funders for this project.

Author Contributions

C.A., Y.L., and M.L. conceptualized and designed the experiments; C.A. collected and analyzed the data and made the figures; M.L. wrote the manuscript with assistance from C.A. and Y.L.

Previous Presentation

These results were previously presented at 2024 Polarization Research Lab Conference, New Orleans, LA, USA.

Preprints

A preprint of this article is published at https://osf.io/v23x5/.

Data Availability

All data and code needed to replicate this analysis are available at: https://osf.io/v23x5/.

References

1 America has a free speech problem. *New York Times*, page 18 March, 2022.
2 Dias N, Druckman J, Levendusky M. Forthcoming. Unraveling a 'cancel culture' dynamic: when and why Americans sanction offensive speech. *J Polit*. https://doi.org/10.2139/ssrn.4235680
3 Smaldino PE. 2022. Models of identity signaling. *Curr Dir Psychol Sci*. 31(3):231–237.
4 Jordan J, Hoffman M, Bloom P, Rand D. 2016. Third-party punishment as a costly signal of trustworthiness. *Nature*. 530(7591): 473–476.
5 Marie A, Petersen MB. 2023. Speech repression and outrage from orthodox activists as attempts at facilitating mobilization and gaining status among allies. *Psychol Inq*. 34(3):192–197.
6 Marwick A. 2021. Morally motivated networked harassment as normative reinforcement. *Soc Media Soc*. 7(2):20563051211021378.
7 Skitka L, Hanson B, Scott MG, Wisneski D. 2021. The psychology of moral conviction. *Annu Rev Psychol*. 72(1):347–366.
8 Grubbs J, Warmke B, Tosi J, James AS, Campbell WK. 2019. Moral grandstanding in public discourse: status-seeking motives as a potential explanatory mechanism in predicting conflict. *PLoS One*. https://doi.org/10.1371/journal.pone.0223749
9 Akerlof G, Kranton R. 2000. Economics and identity. *Q J Econ*. 115(3):715–753.
10 Iannaccone L. 1992. Sacrifice and stigma: reducing free-riding in cults, communes, and other collectives. *J Polit Econ*. 100(2): 271–291.
11 White I, Laird C. 2020. *Steadfast democrats: how social forces shape black political behavior*. Princeton (NJ): Princeton University Press.
12 Lee AH-Y. 2020. How the politicization of everyday activities affects the public sphere: the effects of partisan stereotypes on cross-cutting interactions. *Polit Commun*. 38(5):499–518.
13 Settle J. 2018. *Frenemies: how social media polarizes America*. New York (NY) and Cambridge (UK): Cambridge University Press.
14 Boyd R, Richerson PJ. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol Sociobiol*. 13(3):171–195.
15 Brady W, Crockett MJ, Van Bavel J. 2020. The mad model of moral contagion: the role of motivation, attention and design in the spread of moralized content online. *Perspect Psychol Sci*. 15(4): 978–1010.
16 Van Bavel J, Robertson C, del Rosario K, Rasmussen J, Rathje S. 2024. Social media and morality. *Annu Rev Psychol*. 75:311–340. https://doi.org/10.1146/annurev-psych-022123-110258
17 Brady W, McLoughlin K, Noan T, Crockett MJ. 2021. How social learning amplifies moral outrage expression in online social networks. *Sci Adv*. 7(33). https://doi.org/10.1126/sciadv.abe5641
18 *[dataset], Ahn C, Lelkes Y, Levendusky M. 2024. Data and code for "sanctioning political speech on social media is driven by partisan norms and identity signaling". OSF https://osf.io/v23x5/.
19 Groenendyk E, Kimbrough E, Pickup M. 2023. How norms shape the nature of belief systems in the mass public. *Am J Pol Sci*. 67(3):623–638.
20 Vogels E. 2022. *A growing share of Americans are familiar with "cancel culture"*. Pew Research Center. https://bit.ly/3y4FQP0.
21 Robertson C, del Rosario K, Van Bavel J. 2024. Inside the funhouse mirror: how social media distorts perceptions of norms. *Curr Opin Psychol*. 60:101918.
22 Druckman J, Levendusky M, McLain A. 2018. No need to watch: how the effects of partisan media can spread via interpersonal discussions. *Am J Pol Sci*. 62(1):99–112.