Working Paper **Please do not cite. If you have a question, email me at taegyoon@psu.edu**

Violent Political Rhetoric on Social Media and Affective Polarization

Taegyoon Kim ✉ | Political Science and Social Data Analytics, Penn State University

How does exposure to social media content promoting political violence influence affective polarization? Although existing research reveals that engaging in political communication on social media can be highly consequential for mass political polarization, little is known about how exposure to content promoting political violence shapes the level of affective polarization. I claim that social media posts promoting political violence have heterogeneous effects depending on its partisan source. While exposure to threats from the opposing party leads to affective polarization by triggering anger, exposure to threats from one's own party contributes to affective depolarization by evoking collective shame. Furthermore, I also examine the moderating effect of the social status of those who deliver violence-promoting content. I test my arguments using a two (threat vs. non-threat) × two (in-party vs. out-party) experiment where participants are treated to four different versions of fabricated social media posts on a set of controversial political issues.

1. Introduction

Mass political polarization is one of the most heavily studied topics in the contemporary U.S. politics. Research shows that not only is the population divided along major policy issues but also, perhaps to a grater extent, people increasingly dislike and distrust those who identify with the other political party (Iyengar et al. 2012, Iyengar et al. 2019). Recent works on "affective polarization" demonstrate that over the last several decades the level of affective polarization has been continuously rising: negative feelings toward members of the other party are on the steady increase while feelings toward one's own party remain stably positive. From a normative perspective, affective polarization is associated with a number of inimical consequences, including reduced potential for political deliberation and compromise (Hutchens et al. 2019, MacKuen et al. 2010). Although partisanship has been studied as a benign trait that helps citizens form policy preferences and make vote decisions, both in American and comparative politics (Converse 1964, Lupu 2013, Lupu 2016), increasing animosity between partisan opponents in the contemporary U.S. is even reminiscent of political contexts where rising mass political polarization between partisan opponents encourages conflict and support for non-democratic means for conflict resolution (Lynch et al. 2017). Then, what makes people dislike and distrust others with different partisan affiliation?

In this paper, I pay attention to recent changes in the media environment that take place in the context of increased partisan animosity. In particular, I investigate how political communication on social media that threatens, endorses, or incites political violence against partisan opponents impacts affective polarization. Building on scholarship on media effects, mass political polarization, offline and online political violence, I propose that exposure to promotion of violence has heterogeneous effects for affective polarization depending on its partisan source. That is, political violence promoted by members of the opposing party leads to affective *polarization* by triggering anger on the part of those who are exposed. Being a target of violent threats by the out-party, individuals experience moral anger toward the out-party and thus evaluate the party and its members more negatively. In contrast, political violence promoted by members of one's own party contributes to affective *depolarization* by evoking collective shame. Observing violation of norm of non-violence by the in-party, individuals experience collective shame on behalf of the in-party aggressor who is psychologically linked with them and respond to such shame by affectively distancing themselves from the in-party. Furthermore, I claim that high social status of the aggressor magnifies the main effects because individuals tend to hold stronger expectations for those who are highly visible to the public as well as socially and politically influential (e.g., opinion leaders or politicians) so the latter's violation of norm of non-violence evoke stronger emotional responses. I test the above arguments using a two (presence of violence promotion vs. absence of violence promotion) × two (in-party vs. out-party) experiment where participants are treated to four different versions of fabricated social media posts on a set of controversial political issues.

In this paper, I make contributions to three streams of research in political science. First, I contribute to scholarship on aggressive political communication by highlighting violent political rhetoric on social media. Although recent research on political communication has revealed that (online) political discourse is often uncivil and hateful (Munger 2017a, Munger 2017b, Siegel 2018, Siegel et al. 2019, Gervais 2015, Gervais 2019, Popan et al. 2019, Suhay et al. 2018), we have yet to examine the most aggressive type of rhetoric that threatens, endorses, or incites physical violence against partisan opponents. My work is among the first attempts to explore causal effects of violent political rhetoric on individuals' political attitudes (see also Zeitzoff 2020). Furthermore, I show that the effects of aggressive political communication can be heterogeneous depending on the alignment of partisan affiliation between those who use aggressive rhetoric and those who are exposed to it. Although cutting-edge survey experiments investigate how partisans respond to in-group politicians' use of violent rhetoric (Zeitzoff 2020), no attempt has been made to investigate distinct (or even contrasting) effects depending on the source of such rhetoric. Given empirical findings in political communication that counter-attitudinal exposure on social media is frequent (Bakshy et al. 2015, Barberá 2014) and that cross-cutting such exposure has important consequences (e.g., for issue polarization, see Bail et al. 2018), it is important to understand the effects of counter-attitudinal as well as pro-attitudinal exposure. By examining how individuals' evaluation of partisan opponents respond to violent political rhetoric both from an in-group member and from an out-group member, I shed new light on attitudinal consequences of aggressive political communication.

In addition, I contribute to the literature on mass political polarization by delineating causal mechanisms leading to a specific type of polarization. Scholarship on affective polarization (Iyengar et al. 2012, Iyengar et al. 2019) has shown that negative views of political opponents are distinct from disagreement over political/policy issues (Abramowitz and Saunders 2008, Fiorina and Abrams 2008). Similarly, scholars have also shown that perceived polarization (the level of political polarization perceived by individuals) differs from both affective and issue polarization.[1] By specifically focusing on affective polarization as opposed to the other forms of mass political polarization (i.e., issue polarization or perceived polarization), I theorize and demonstrate that emotional responses to violent political content matter for evaluations of partisan opponents and also call for further investigation into whether such mechanisms can travel to the other forms of mass political polarization.

Lastly, I contribute to the literature on the effects of group-based emotions on political attitude and behavior (Chudy et al. 2019, Marcus 1988, Marcus and MacKuen 1993, Wayne 2019).[2] Although much work in political science exists examining the role of group-based emotions for ethnic (e.g., Chudy et al. 2019) and international relations (e.g., Wayne 2019), little is known about their role for inter-party relations. As partisanship and ideology align with each other and they are increasingly a crucial part of individuals' social identity (Mason 2015), it is very important to understand how individuals process an external stimulus (e.g., violent political rhetoric) on behalf on the party they identify with. By highlighting the role of collective shame as well as anger, I shed new light on how group-based emotional responses shape inter-party relations in context of aggressive political communication.

2. Related Work

The Rise of Affective Polarization

Mass political polarization is one of the most hotly debated topics in U.S. politics. Focused on ideology, it has most commonly been defined and studied in terms of the divergence between Republicans and Democrat's attitudes toward major policy issues. In contrast to the clear political significance of such divergence, however, empirical evidence for the presence of issue polarization (or attitude or ideological polarization) among the mass is less clear. While there is much evidence that political elites are polarized over policy issues (Jacobson 2005, Poole and Rosenthal 2001, Ladewig 2010), the extent of polarization among the mass is ambiguous. Some scholars have found that voters are also becoming more ideologically polarized, while others claim that people are still much more moderate than their leaders (Abramowitz and Saunders 2008, Fiorina and Abrams 2008)).

More recently, scholarship in various subfields of political science has highlighted another, equally important, type of mass political polarization. Highlighting increasing animosity between the two major parties, a large group of scholars started to investigate what is called "affective polarization." Affective polarization is defined as the degree to which citizens dislike and distrust others identified with the other party (Iyengar et al. 2019). In U.S., the last several decades have seen a consistent upward trend in negative feelings toward members of the other party (Iyengar et al. 2012) while feelings toward one's own party have remained stably positive, together leading to a steady increase

[1] For ideological polarization, see Abramowitz and Saunders (2008), Barberá (2014), and Barberá et al. (2015), Fiorina and Abrams (2008). For affective polarization, see Druckman et al. (2019), Iyengar et al. (2012), Iyengar et al. (2019), and Popan et al. (2019). For perceived polarization, see Enders and Armaly (2019), Lelkes (2016), and Settle (2018).

[2] For a comprehensive review of research on the role of emotions in political science, see Groenendyk (2011).

in affective polarization. Affective polarization is associated with a number of inimical consequences. Hutchens et al. (2019) shows that affective polarization and exclusive partisan homophily in political discussion form a reinforcing spiral over time, implying creation of partisan echo-chambers filled with cross-partisan hatred. Even when citizens engage in discussions with those who are from the other party, research predicts that cross-partisan dialogue will be anti-deliberative. For instance, MacKuen et al. (2010) shows that anger, a definitive feature of affective polarization, compromises potential for compromise and reinforces prior attitudes.

Partisan Media, Social Media, and Affective Polarization

Seeking to explain increasing level of affective polarization, a stream of works point out that (hyper) partisan news outlets make partisanship salient thereby inducing increased affective polarization (Lau et al. 2017, Lelkes et al. 2017). Highlighting the shift in the media ecosystem from the largely homogeneous broadcast news era in the 1980s to the high-choice environment that emerged with the introduction of cable televisions, these works claim that individuals are now able to self-select into pro-attitudinal partisan news outlets where their in-group identity is primed thereby evaluating their party more positively and the other party more negatively (Levendusky 2013). In addition, depiction of the out-party and its members in negative light is a common practice among many partisan news outlets, which in turn can encourage inter-party hostility (Berry and Sobieraj 2013). Nevertheless, evidence for the polarizing effects of partisan news outlets is far from clear. First, Arceneaux and Johnson (2013) point out that those who are already highly polarized self-select into pro-attitudinal partisan news outlets and thus the aggregate effects of the exposure is minimal. Contradicting the polarizing effects from another perspective, Gentzkow and Shapiro (2011) argue that even when partisan news outlets have polarizing effects individuals generally prefer identity-consistent neural content.

Similar to literature emphasizing the polarizing effects of partisan news outlets on affective polarization, the prevalent narrative in literature focused on social media is that social media users associate only with those who share their partisan preferences, thus forming an “echo-chamber” where the consumption of pro-attitudinal information from like-minded individuals reinforces prior attitudes while exposure to disagreements seldom takes place. Focusing on Twitter in U.S., Colleoni et al. (2014) claims that there is a significant level of political homophily in Twitter following networks. Similarly, Conover et al. (2011) also reports that Twitter retweet networks on political topics in U.S. display two distinct ideological clusters of the conservative and liberal. Discussing the consequences of lack of cross-cutting exposure, Sunstein (2018) argues that social media will result in an increasingly segregated society where citizens are polarized along partisan lines, undermining the potential for political compromise.

However, there is also mounting evidence against the aforementioned “echo-chamber" arguments. Indeed, there are increasing acknowledgements that social media also expose individuals to counter-attitudinal content and may even lead to political depolarization. Diffusion tools built in social media platforms may expose individuals to content that is not actively or explicitly sought out (Eady et al. 2019). Such content is most likely from weak social ties and tend to be counter-attitudinal. Consequently, despite the homophilic nature of personal networks (McPherson et al. 2001), social media leads to exposure to a wider range of political opinions than one would normally encounter offline (Barberá 2014, Bakshy et al. 2015). Indeed, Barberá (2014) demonstrates that social media users embedded in ideologically heterogeneous networks are more likely to moderate their political ideology over time. Focusing on Egypt, Siegel et al. (2019) claims that Twitter users embedded in ideologically heterogeneous (Islamic users and secular users) networks become more politically tolerant to outgroup members over time. Taken together, they not only provide further evidence that there is a significant amount of ideological heterogeneity on social media but also suggest that exposure to such heterogeneity can be highly consequential for political moderation.

Although I do not to take on this ongoing debates, three general points about the aforementioned works on the polarizing effects of social media need to be made. First, these works richly demonstrate that, despite general tendency for homophily and selective exposure, individuals are exposed to substantial amount of counter-attitudinal information that they would not in offline - partisan media in particular - settings. This is important because, as demonstrated in the subsequent sections, attitude-consistency of political information (or whether the political information comes from an in-party or out-party source) can lead to heterogeneous effects. Second, although existing works show that political information on social media can be either from an in-party or from an out-party source, it is not the only theoretically important dimension. The next section shows that the degree of violence (or aggressiveness in general) in political content on social media is another crucial factor that shapes the influence of social media on affective polarization. Last, recent studies on the polarizing effects of social media are largely focused on issue polarization rather than affective polarization. This void calls for further research specifically dedicated to the effects of social media on affective polarization.

The Effects of Violent Political Communication on Social Media

Contradicting the optimistic depiction of social media as a place for political learning and deliberation (Dimitrova et al. 2014, Halpern and Gibbs 2013), the current climate of political communication on social media is tainted by aggressive partisan expressions. The reduced gate-keeping power of traditional media outlets and online anonymity that came along make it easier for social media users to express uncivil and hateful opinions targeted at people of different race, gender, and partisan affiliation (Berry and Sobieraj 2013, Munger 2017a, Munger 2017b). According to a survey, 93% of the survey respondents perceive incivility as a major problem in the country and 63% believe that social media contributes to incivility. The survey also reports that the number of encounters with uncivil online content has grown (from 4.4 encounters per week in 2013 to 5.5 in 2019) (Shandwick 2019). Recent research on political communication also suggests there is a significant amount of incivility (Coe et al. 2014, Middaugh et al. 2017, Rowe 2015, Siegel et al. 2019, Sobieraj and Berry 2011). In addition to the upward trend toward incivility, research also finds that minority populations are targeted with hate speech for expressing different views (Kennedy and Taylor 2010, Matamoros-Fernández 2017).[3].

Despite such effort, however, scholarship on political communication have yet to pay sufficient attention to violent rhetoric that threats, endorses, and incites political violence

[3] For a comprehensive review of studies on online hate speech, see Siegel (2018)

on social media. As demonstrated in the first essay of this dissertation, social media users use violent rhetoric to express their extreme political views as well as threaten, endorse, and incite political violence. Although such messages are small in number, they can spread to a large number of benign users through multiple chains of communication networks. Fortunately, recent works in both political and non-political domains suggest that exposure to online violence is associated with a host of harmful consequences for individuals. Works on violent extremism suggest that exposure to violent extremist content on social media is robustly associated with individuals' propensity to commit political violence (Pauwels and Schils 2016). In addition, works on cyber violence in non-political domains suggest that cyber bullying (repetitive aggressive online behavior that involves power imbalance between the perpetrator and the victim) can cause a number of negative social and psychological consequences, including poor school performance, negative self-esteem, anxiety, depression, isolation, loneliness, stress, and even suicidality (Ang 2015, Baek and Bullock 2014, Baldry et al. 2015, Cassidy et al. 2013, Garett et al. 2016).

3. Theory

How does exposure to social media posts threatening, endorsing, and inciting political violence influence affective polarization? As demonstrated by rich scholarship on online political communication, social media users experience counter-attitudinal exposure to a substantial degree (Bakshy et al. 2015, Barberá 2014, Eady et al. 2019) and such exposure can be often aggressive (Munger 2017a, Sobieraj and Berry 2011). Accordingly, it is crucial to consider potentially heterogeneous effects of violence-promoting posts on affective polarization depending on whether the violent content comes from an member of the in-party or the out-party. I suggest that the effects of exposure to social media posts promoting political violence can be heterogeneous because different sources of threats evoke distinct emotional responses . While exposure to threats from the opposing party leads to affective polarization by triggering either fear or anger, exposure to threats from one's own party contributes to affective depolarization by evoking vicarious shame.

Threats from the Out-party

First of all, I claim that promotion of violence by an out-party member targeting the in-party leads to affective polarization primarily by exacerbating negative feelings such as fear or anger toward the out-party. Although the relationship between threats and group-relations has heavily been studied in the field of ethnic conflict and international relations, the aggressive political climate in U.S. between partisan opponents and increasing salience of partisanship as a social identity make inter-partisan relationships on social media highly analogous to the inter-group relationships studied in the aforementioned fields. From a comparative perspective, physical political violence in U.S. is rather limited (Kishi and Carboni 2019). Nevertheless, violent partisanship (Kalmoe and Mason 2018), coupled with unfiltered extreme partisan expressions exacerbated by the absence of gate-keeping role of traditional media actors, makes political communication on social media prone to group-based conflict and insights from scholarship on inter-group conflict highly relevant.

According to social identity theory, in-group identification is such an essential part of self that provides individuals with acceptance and belonging as well as guides their social behavior to the extent that the mere presence of inter-group antagonism is its natural corollary (Tajfel et al. 1979). Not only so, a long tradition of research on inter-group relations demonstrates that, across many contexts, threats coming from the out-group have a host of negative consequences for individuals' attitudes toward the out-group. For instance, works on terrorism show that individuals generally deal with threats by adopting aggressive attitudes toward the out-group. Individuals exposed to terrorism are more likely to support for hostile foreign policies (Getmansky and Zeitzoff 2014), exhibit reduced political tolerance (Shamir and Sagiv-Schifter 2006), and exclusionist attitudes against ethnic minorities (Canetti-Nisim et al. 2009).

Focusing on emotional responses to threats, works on intergroup threat theories clearly show that threatened individuals experience negative emotions such fear, anger, or contempt (Stephan et al. 2008, Renfro et al. 2006, Mackie et al. 2000). A similar stream of works further suggest that anger is common when the entire in-group is threatened while fear is dominant when threats are targeted at an individual (Cottrell and Neuberg 2005). In the former, those who threatened experience fear out of concerns about their personal security or self-image whereas, in the latter, anger is often evoked due to (potential) loss of the in-group's resources or reputation. In the same vein, recent research on individual responses to terrorism suggests that predominant responses to terrorism is anger rather than fear and that individuals who feel the least personally threatened but feel morally outrageous adopt hostile foreign policy preferences in the wake of terrorist attacks (Wayne 2019).

Social media posts promoting violence against the opposing party are targeted typically at the out-party or its members as a whole. In addition, when such threats are targeted at specific individuals, the target is often high-profile politicians rather than ordinary citizens on social media. Therefore, it is possible to expect that predominant emotions evoked in response to social media posts promoting violence against the in-party will likely be anger. Regardless of the specificity of the target, however, it is abundantly clear that the consequences of threats of political violence for affective evaluation toward the out-party will be uniformly negative. Therefore, I expect that exposure to promotion of violence targeted at the in-group will lead to affective polarization by evoking anger and deteriorating feelings towards the out-party.

Threats from the In-party

As opposed to promotion of violence from an out-party member, I claim that violence promoted by an in-party member have depolarizing effects by evoking collective partisan shame. According to psychological theories, shame is a powerful emotion of self-condemnation that regulates many social interactions in situations where one has violated a social norm (Eisenberg 2000, Keltner and Harker 1998, Tangney and Fischer 1995). Accordingly, I define partisan collective shame as shame that partisans experience on behalf of their in-party members' actions deemed normatively undesirable. A vast majority of citizens on social media rarely engage in actions that promote political violence. To a varying degree, however, they are exposed to extremist in-group members' promotion of violence against the out-group. Even though they themselves are not directly involved in such a violent act, they can experience vicarious shame on behalf of their in-party because they are psychologically linked with the extremist in-group member by a shared group membership.[4]

I suggest that those who experience collective shame respond to such a negative experience by emotionally detaching oneself from the self-identified in-group. Although they will not go on to forsake their partisanship, I expect that emotional detachment from the in-group will be a spontaneous or unconscious response resulting from desire to maintain positive self-image. At the same time, it is also possible that experience of collective partisan shame leads to more positive evaluation of the out-group. Collective shame is known to motivate reparative and compensatory attitudes in various contexts (Brown et al. 2008, Chudy et al. 2019, Swim and Miller 1999). I thus suggest that reduced negative evaluation of the out-group will be observed after exposure to in-group promotion of violence. All in all, the consequence of exposure to in-group members' promotion of violence is reduced affective polarization.

Negative emotions experienced on behalf of the in-group have often been studied from an cross-generational perspective such as one's county's history of colonization (Doosje et al. 1998) or one's ethnic racial group's past human right violations againt another racial group (Chudy et al. 2019, Imhoff et al. 2012). To my knowledge, however, the literature has yet to examine how collective shame can be manifest in contemporaneous inter-partisan context. Although U.S. citizens are increasingly hostile to those who do not share their partisan affiliation, only a small fraction of them openly endorse or take pleasure form use of violence against partisan opponents. For instance, using nationally-representative surveys, Kalmoe and Mason (2018) shows that around 20 percent of the population believes that it is occasionally acceptable to send threatening messages to public officials and around 9 percent thinks that violence could be acceptable if their partisan opponent won the 2020 Presidential Election. Although these statistics are surprisingly high, they still mean that a vast majority of the population is aversive to use of political violence. Indeed, this is in part why individuals exposed to promotion of violence by an out-party member experience moral outrage. By the same token, I expect that, to the extent that one shares the norm of non-violence in domestic political competition, exposure to in-group members' promotion of political violence will lead to affective depolarization.

Social Status of Aggressor

I further argue that the heterogeneous treatment effects described above will be moderated by the social status of the aggressor. The core assumption of the treatment effects hypothesized above is that individuals perceive acts of promoting violence for political purposes as breaking norm of non-violence. I argue that the social status of the aggressor varies the intensity of norm of non-violence in online political communications and thus moderate the main effects. Specifically, the social status of the perpetrator affects the level of norm of non-violence *expected of* those who deliver political content.

One important characteristic shared by major social media platforms such as Facebook,

[4] In a similar vein, cutting-edge works on the effects of political incivility on affective polarization also suggest that in-party incivility has depolarizing effects in context of partisan media and politicians' communication on social media (Druckman et al. 2019, Gervais 2019). However, they do not discuss why uncivil rhetoric leads to affective depolarization.

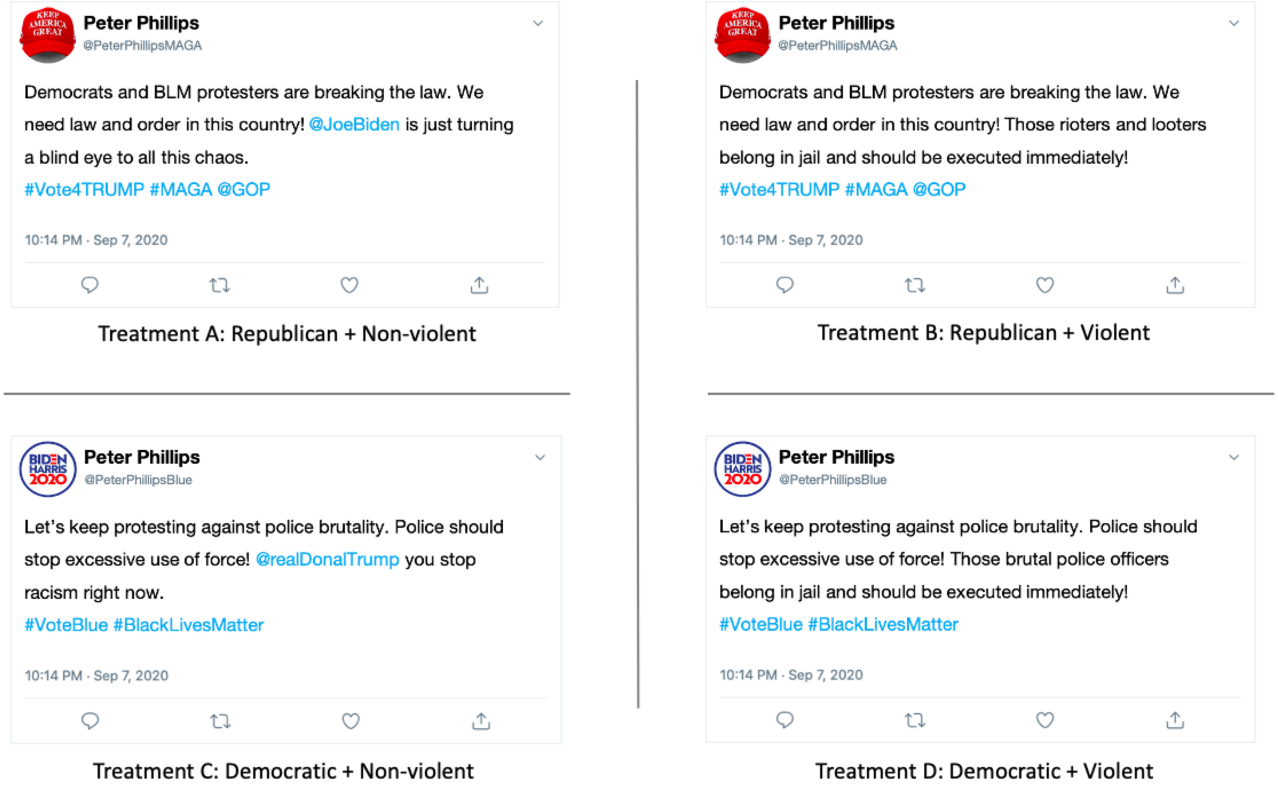
Twitter, and Instagram is that ordinary citizens are on the same communication space with more socially and politically influential actors such as opinion leaders and politicians. Therefore, the social status of those who deliver violence-promoting can either be ordinary partisans or be "social referents" who are well-connected and socially visible individuals and whose behavior serve as sources of normative information (Paluck and Shepherd 2012, Paluck et al. 2016). Here, I claim that partisan social referents' (e.g. slanted opinion leaders or politicians) failure to abide by norm of non-violence can have important consequences for affective polarization because individuals expect high level of norm abidance from social referents. In addition, because social referents are highly visible on social media, individuals normative expectations from them can be even higher. Therefore, both the polarizing effect of violence promotion from an out-group member and the depolarizing effect of violence promotion from an in-group member will be stronger by evoking greater anger and shame, respectively.

4. Experiment

Treatments

Figure 1 shows four example Tweets that will be used as treatments.[5] Note that all of the Tweets are synthetic but are designed to reflect recent partisan tension around a salient political issue. That is, each of the Tweets expresses an opinion about recent protests against racism and police violence. Treatments A and B, on the top, are presented as written by the same Republican user but they differ in their use of violent rhetoric. The author in Treatment A is condemning Democrats and Black Lives Matter protesters for violence during the protests. It also criticizes for the Democratic presidential candidate Joe Biden for conniving such violence. Treatment B differs from Treatment A only in use of violence rhetoric: "Those rioters and looters belong in jail and should be executed immediately." Other aspects of the two Tweets are kept identical, including the profile picture, the author's name, the account name, the hashtags, the mentions, and the time of publication. Treatments C and D, at the bottom, are a Democratic mirror image. The author in Treatment C criticizes Republicans and Donald Trump, the incumbent president and Republican presidential candidate, for conniving or justifying police violence. Similar to the contrast between Treatments A and B, Treatments C and D only differ in use of violent rhetoric. While Treatment C is simply calling for Trump to stop racism, Treatment D is calling for violence against police officers.

[5] To test the heterogeneous treatment effects depending on the social status of the aggressor, there will be another set of four Tweets that only vary in the author's social status.



Peter Phillips
@PeterPhillipsMAGA

Democrats and BLM protesters are breaking the law. We need law and order in this country! @JoeBiden is just turning a blind eye to all this chaos.
#Vote4TRUMP #MAGA @GOP

10:14 PM · Sep 7, 2020

Treatment A: Republican + Non-violent

Peter Phillips
@PeterPhillipsMAGA

Democrats and BLM protesters are breaking the law. We need law and order in this country! Those rioters and looters belong in jail and should be executed immediately!
#Vote4TRUMP #MAGA @GOP

10:14 PM · Sep 7, 2020

Treatment B: Republican + Violent

Peter Phillips
@PeterPhillipsBlue

Let's keep protesting against police brutality. Police should stop excessive use of force! @realDonalTrump you stop racism right now.
#VoteBlue #BlackLivesMatter

10:14 PM · Sep 7, 2020

Treatment C: Democratic + Non-violent

Peter Phillips
@PeterPhillipsBlue

Let's keep protesting against police brutality. Police should stop excessive use of force! Those brutal police officers belong in jail and should be executed immediately!
#VoteBlue #BlackLivesMatter

10:14 PM · Sep 7, 2020

Treatment D: Democratic + Violent

Figure 1: *Example Treatments*

Ethical Considerations

The central theoretical interest, threats of political violence, might pose risks to the participants when used as an experimental stimulus. However, absolute minimization of risks is not the only consideration for the ethics for experimental research. Given the salience of threats of violence in online political communication and its potentially detrimental effects on peaceful resolution of inter-party conflict, understanding of its consequences through rigorous experimental research is urgently needed for the benefits of the broader society, particularly for the maintenance of the current U.S. democracy. Nevertheless, given findings in the previous research that similar rhetoric, violent political metaphors, has negative consequences (see Kalmoe 2014 for increased tolerance of violence, Kalmoe et al. 2018 for issue polarization, Kalmoe 2019 for voter (de)mobilization), it is important to carefully consider potential costs imposed on the participants and the broader society and come up with ways to minimize them.

First, I plan to inform potential participants of the general goal and content of the experiment with a particular focus on the fact that it is about fierce interpartisan tension revealed in social media communication. This is in line with the principle of respecting potential participants' autonomy (as in the 1979 Belmont Report). Although this might let participants who are more tolerant of aggression self-select into the experiment, I plan to measure the level of relevant psychological characteristics (e.g. aggressive personality) to make sure the distribution on such characteristics are not overly skewed.

Second, exposure to violent threats can be emotionally and cognitively disturbing even if they are not directly targeted at the participants. Certainly, negative evaluation of the out-partisans — the outcome of interest — can be one of potential risks. Although some level of antipathy between opposing partisans is inevitable and even the essence of party politics in many contexts, there are rising concerns about the current level of affective polarization in the U.S. However, it is far from straightforward whether this experiment will deteriorate the current level of affective polarization for two reasons. At the very least, the theorized effect of exposure to threats of violence by a co-partisan will actually lower the level of affective polarization. In addition, the size of the participants (N = 1,500-2,000) is unlikely to pose any serious threat to the inter-party relations. This is in stark contrast to some online field experiment where the emotion/behavior of millions of people were manipulated and could have led to huge-scale effects in aggregate terms (for instance, see Kramer et al. 2014).

Third, although threats of violence can be disturbing, violent political communication on social media, including threats of violence, is increasingly common in social media platforms.[6] Recent political turmoil around President Trump's threatening Tweets against Black Lives Matter protesters and the ensuing hostile communication on Twitter exemplify normalization of violent political rhetoric as daily partisan discourse to a certain extent. Therefore, I argue that exposure to such rhetoric is part of within the normal expectations of user experience on social media. Indeed, existing political science works have exposed participants to various types of violent political rhetoric, both violent political metaphors and threats of violence (see Kalmoe 2014, Kalmoe et al. 2018, and Kalmoe 2019 for violent political metaphors and see Zeitzoff 2020 for threats of violence). Similarly, there is an extensive body of experimental literature on the effects of violent media on real-world aggression where participants are exposed to various forms of media violence, including lyrics, movies, and video games (Arriaga et al. 2015, Bender et al. 2018, Gentile et al. 2017, Giancola and Parrott 2008, Plante and Anderson 2017). Although the intensity of violent treatment is not without limits in this literature as well, I claim that the particular type of violent content highlighted in this study (shown in Figure 1) is no more risky than the treatments used in the existing literature on other violent content.[7]

Finally, after the experiment, I plan to debrief the participants. Since the participants are not informed of the fact that the treatment Tweets are synthetic prior to the experiment, it is a form of deception. This choice is inevitable for experimental validity because, if participant were informed of the deception, they would react in a very different manner. However, use of synthetic social media posts is not just common among studies on online communication, both in lab experiments (Popan et al. 2019, Suhay et al. 2018) and field experiments (Gallego et al. 2019, Munger 2017b, Munger 2017a), but also more ethically justifiable than using synthetic accounts that interact with human subjects (Gallego et al. 2019, Mønsted et al. 2017, Munger 2017b, Munger 2017a). The potential risk is, although this particular form of deception is fairly mild, this might lead the participants to be misinformed about the reality. For instance, they might overestimate the degree to which social media users use rhetoric containing threats of violence, the degree to which Twitter is used for such rhetoric, and so on. In addition to informing them that the Tweets are

[6] See reports on threats of violence against *politicians* as well as *citizens*.

[7] Also, for treatment synthetic Tweets, I plan to avoid direct expressions that evoke images of bodily harm.

synthetic, I plan to provide them informative explanation of the rationale for the design of the experiment, the hypotheses tested, and the methods used, and ask for and answer further questions. I expect this will minimize the potential risk posed by concealing the fact that the Tweets used as the treatment are synthetic.

References

Abramowitz, A. I. and K. L. Saunders (2008). Is polarization a myth? *The Journal of Politics 70*(2), 542–555.

Ang, R. P. (2015). Adolescent cyberbullying: A review of characteristics, prevention and intervention strategies. *Aggression and violent behavior 25*, 35–42.

Arceneaux, K. and M. Johnson (2013). *Changing minds or changing channels?: Partisan news in an age of choice*. University of Chicago Press.

Arriaga, P., J. Adrião, F. Madeira, I. Cavaleiro, A. Maia e Silva, I. Barahona, and F. Esteves (2015). A "dry eye" for victims of violence: Effects of playing a violent video game on pupillary dilation to victims and on aggressive behavior. *Psychology of Violence 5*(2), 199.

Baek, J. and L. M. Bullock (2014). Cyberbullying: a cross-cultural perspective. *Emotional and behavioural difficulties 19*(2), 226–238.

Bail, C. A., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences 115*(37), 9216–9221.

Bakshy, E., S. Messing, and L. A. Adamic (2015). Exposure to ideologically diverse news and opinion on facebook. *Science 348*(6239), 1130–1132.

Baldry, A. C., D. P. Farrington, and A. Sorrentino (2015). "am i at risk of cyberbullying"? a narrative review and conceptual framework for research on risk of cyberbullying and cybervictimization: The risk and needs assessment approach. *Aggression and Violent Behavior 23*, 36–51.

Barberá, P. (2014). How social media reduces mass political polarization. evidence from germany, spain, and the us. *Job Market Paper, New York University 46*.

Barberá, P., J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science 26*(10), 1531–1542.

Bender, P. K., C. Plante, and D. A. Gentile (2018). The effects of violent media content on aggression. *Current opinion in psychology 19*, 104–108.

Berry, J. M. and S. Sobieraj (2013). *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Brown, R., R. González, H. Zagefka, J. Manzi, and S. Čehajić (2008). Nuestra culpa: collective guilt and shame as predictors of reparation for historical wrongdoing. *Journal of personality and social psychology 94*(1), 75.

Canetti-Nisim, D., E. Halperin, K. Sharvit, and S. E. Hobfoll (2009). A new stress-based model of political extremism: Personal exposure to terrorism, psychological distress, and exclusionist political attitudes. *Journal of Conflict Resolution 53*(3), 363–389.

Cassidy, W., C. Faucher, and M. Jackson (2013). Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School psychology international 34*(6), 575–612.

Chudy, J., S. Piston, and J. Shipper (2019). Guilt by association: White collective guilt in american politics. *The Journal of Politics 81*(3), 968–981.

Coe, K., K. Kenski, and S. A. Rains (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication 64*(4), 658–679.

Colleoni, E., A. Rozza, and A. Arvidsson (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication 64*(2), 317–332.

Conover, M. D., J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini (2011). Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Converse, P. E. (1964). The nature of belief systems in mass publics. ideology and discontent. *Ideology and Discontent*, 206–261.

Cottrell, C. A. and S. L. Neuberg (2005). Different emotional reactions to different groups: a sociofunctional threat-based approach to" prejudice". *Journal of personality and social psychology 88*(5), 770.

Dimitrova, D. V., A. Shehata, J. Strömbäck, and L. W. Nord (2014). The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data. *Communication research 41*(1), 95–118.

Doosje, B., N. R. Branscombe, R. Spears, and A. S. Manstead (1998). Guilty by association: When one's group has a negative history. *Journal of personality and social psychology 75*(4), 872.

Druckman, J. N., S. Gubitz, A. M. Lloyd, and M. S. Levendusky (2019). How incivility on partisan media (de) polarizes the electorate. *The Journal of Politics 81*(1), 291–295.

Eady, G., J. Nagler, A. Guess, J. Zilinsky, and J. A. Tucker (2019). How many people live in political bubbles on social media? evidence from linked survey and twitter data. *SAGE Open 9*(1), 2158244019832705.

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual review of psychology 51*(1), 665–697.

Enders, A. M. and M. T. Armaly (2019). The differential effects of actual and perceived polarization. *Political Behavior 41*(3), 815–839.

Fiorina, M. P. and S. J. Abrams (2008). Political polarization in the american public. *Annu. Rev. Polit. Sci. 11*, 563–588.

Gallego, J., J. D. Martínez, K. Munger, and M. Vásquez-Cortés (2019). Tweeting for peace: Experimental evidence from the 2016 colombian plebiscite. *Electoral Studies 62*, 102072.

Garett, R., L. R. Lord, and S. D. Young (2016). Associations between social media and cyberbullying: a review of the literature. *Mhealth 2*.

Gentile, D. A., P. K. Bender, and C. A. Anderson (2017). Violent video game effects on salivary cortisol, arousal, and aggressive thoughts in children. *Computers in Human Behavior 70*, 39–43.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*(4), 1799–1839.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics 12*(2), 167–185.

Gervais, B. T. (2019). Rousing the partisan combatant: Elite incivility, anger, and antideliberative attitudes. *Political Psychology 40*(3), 637–655.

Getmansky, A. and T. Zeitzoff (2014). Terrorism and voting: The effect of rocket threat on voting in israeli elections. *American Political Science Review 108*(3), 588–604.

Giancola, P. R. and D. J. Parrott (2008). Further evidence for the validity of the taylor aggression paradigm. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression 34*(2), 214–229.

Groenendyk, E. (2011). Current emotion research in political science: How emotions help democracy overcome its collective action problem. *Emotion Review 3*(4), 455–463.

Halpern, D. and J. Gibbs (2013). Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in Human Behavior 29*(3), 1159–1168.

Hutchens, M. J., J. D. Hmielowski, and M. A. Beam (2019). Reinforcing spirals of political discussion and affective polarization. *Communication Monographs 86*(3), 357–376.

Imhoff, R., M. Bilewicz, and H.-P. Erb (2012). Collective regret versus collective guilt: Different emotional reactions to historical atrocities. *European Journal of Social Psychology 42*(6), 729–742.

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science 22*, 129–146.

Iyengar, S., G. Sood, and Y. Lelkes (2012). Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly 76*(3), 405–431.

Jacobson, G. C. (2005). Polarized politics and the 2004 congressional and presidential elections. *Political Science Quarterly 120*(2), 199–218.

Kalmoe, N. P. (2014). Fueling the fire: Violent metaphors, trait aggression, and support for political violence. *Political Communication 31*(4), 545–563.

Kalmoe, N. P. (2019). Mobilizing voters with aggressive metaphors. *Political Science Research and Methods 7*(3), 411–429.

Kalmoe, N. P., J. R. Gubler, and D. A. Wood (2018). Toward conflict or compromise? how violent metaphors polarize partisan issue attitudes. *Political Communication 35*(3), 333–352.

Kalmoe, N. P. and L. Mason (2018). Lethal mass partisanship: Prevalence, correlates, and electoral contingencies. In *American Political Science Association Conference*.

Keltner, D. and L. Harker (1998). The forms and functions of the nonverbal signal of shame.

Kennedy, M. A. and M. A. Taylor (2010). Online harassment and victimization of college students. *Justice Policy Journal 7*(1), 1–21.

Kramer, A. D., J. E. Guillory, and J. T. Hancock (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences 111*(24), 8788–8790.

Ladewig, J. W. (2010). Ideological polarization and the vanishing of marginals: Retrospective roll-call voting in the us congress. *The Journal of Politics 72*(2), 499–512.

Lau, R. R., D. J. Andersen, T. M. Ditonto, M. S. Kleinberg, and D. P. Redlawsk (2017). Effect of media environment diversity and advertising tone on information search, selective exposure, and affective polarization. *Political Behavior 39*(1), 231–255.

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly 80*(S1), 392–410.

Lelkes, Y., G. Sood, and S. Iyengar (2017). The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science 61*(1), 5–20.

Levendusky, M. (2013). *How partisan media polarize America*. University of Chicago Press.

Lupu, N. (2013). Party brands and partisanship: Theory with evidence from a survey experiment in argentina. *American Journal of Political Science 57*(1), 49–64.

Lupu, N. (2016). *Party brands in crisis: partisanship, brand dilution, and the breakdown of political parties in Latin America*. Cambridge University Press.

Lynch, M., D. Freelon, and S. Aday (2017). Online clustering, fear and uncertainty in egypt's transition. *Democratization 24*(6), 1159–1177.

Mackie, D. M., T. Devos, and E. R. Smith (2000). Intergroup emotions: explaining offensive action tendencies in an intergroup context. *Journal of personality and social psychology 79*(4), 602.

MacKuen, M., J. Wolak, L. Keele, and G. E. Marcus (2010). Civic engagements: Resolute partisanship or reflective deliberation. *American Journal of Political Science 54*(2), 440–458.

Marcus, G. E. (1988). The structure of emotional response: 1984 presidential candidates. *American Political Science Review 82*(3), 737–761.

Marcus, G. E. and M. B. MacKuen (1993). Anxiety, enthusiasm, and the vote: The emotional underpinnings of learning and involvement during presidential campaigns. *American Political Science Review 87*(3), 672–685.

Mason, L. (2015). "i disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science 59*(1), 128–145.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an australian race-based controversy on twitter, facebook and youtube. *Information, Communication & Society 20*(6), 930–946.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology 27*(1), 415–444.

Middaugh, E., B. Bowyer, and J. Kahne (2017). U suk! participatory media and youth experiences with political discourse. *Youth & Society 49*(7), 902–922.

Mønsted, B., P. Sapieżyński, E. Ferrara, and S. Lehmann (2017). Evidence of complex contagion of information in social media: An experiment using twitter bots. *PloS one 12*(9), e0184148.

Munger, K. (2017a). Experimentally reducing partisan incivility on twitter. *Unpublished working paper. Available at: https://kmunger. github. io/pdfs/jmp. pdf*.

Munger, K. (2017b). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior 39*(3), 629–649.

Paluck, E. L. and H. Shepherd (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of personality and social psychology 103*(6), 899.

Paluck, E. L., H. Shepherd, and P. M. Aronow (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences 113*(3), 566–571.

Pauwels, L. and N. Schils (2016). Differential online exposure to extremist content and political violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence 28*(1), 1–29.

Plante, C. and C. A. Anderson (2017). Media, violence, aggression, and antisocial behavior: Is the link causal? *The Wiley Handbook of Violence and Aggression*, 1–12.

Poole, K. T. and H. Rosenthal (2001). D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting. *Legislative Studies Quarterly*, 5–29.

Popan, J. R., L. Coursey, J. Acosta, and J. Kenworthy (2019). Testing the effects of incivility during internet political discussion on perceptions of rational argument and evaluations of a political outgroup. *Computers in Human Behavior 96*, 123–132.

Renfro, C. L., A. Duran, W. G. Stephan, and D. L. Clason (2006). The role of threat in attitudes toward affirmative action and its beneficiaries 1. *Journal of Applied Social Psychology 36*(1), 41–74.

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, communication & society 18*(2), 121–138.

Settle, J. E. (2018). *Frenemies: How social media polarizes America*. Cambridge University Press.

Shamir, M. and T. Sagiv-Schifter (2006). Conflict, identity, and tolerance: Israel in the al-aqsa intifada. *Political Psychology 27*(4), 569–595.

Shandwick, W. (2019). Civility in america 2019: Solutions for tomorrow.

Siegel, A., E. Nikitin, P. Barberá, J. Sterling, B. Pullen, R. Bonneau, J. Nagler, and J. A. Tucker (2019). Trumping hate on twitter? online hate speech in the 2016 us election campaign and its aftermath.

Siegel, A., J. Tucker, J. Nagler, and R. Bonneau (2019). Tweeting beyond tahrir: Ideological diversity and political tolerance in egyptian twitter networks. *Unpublished working paper, New York University*.

Siegel, A. A. (2018). Online hate speech.

Sobieraj, S. and J. M. Berry (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication 28*(1), 19–41.

Stephan, W. G., C. Renfro, and M. D. Davis (2008). The role of threat in intergroup relations.

Suhay, E., E. Bello-Pardo, and B. Maurer (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics 23*(1), 95–115.

Sunstein, C. R. (2018). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Swim, J. K. and D. L. Miller (1999). White guilt: Its antecedents and consequences for attitudes toward affirmative action. *Personality and Social Psychology Bulletin 25*(4), 500–514.

Tajfel, H., J. C. Turner, W. G. Austin, and S. Worchel (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56–65.

Tangney, J. P. E. and K. W. Fischer (1995). Self-conscious emotions: The psychology of shame, guilt, embarrassment, and pride. In *The idea for this volume grew out of 2 pivotal conferences. The 1st conference, on emotion and cognition in development, was held in Winter Park, CO, Sum 1985. The 2nd conference, on shame and other self-conscious emotions, was held in Asilomar, CA, Dec 1988.* Guilford Press.

Wayne, C. (2019). *Risk or Retribution: The Micro-foundations of State Responses to Terror*. Ph. D. thesis.

Zeitzoff, T. (2020). The nasty style: Why politicians use violent rhetoric. *Unpublished working paper. Available at: https://www.zeitzoff.com/uploads/2/2/4/1/22413724/zeitzoff*$_{n}astystyle_{v}iolentrhetoric_{d}raft_{f}eb2020.pdf$.