Asymmetric Affective Polarization in Elite and Public Discourse: Event-Time Evidence from the Black Lives Matter Uprising

Sylvia Chang Dou[0009−0000−6215−9963]

University of Chicago, Chicago IL 60637, USA
douchang@uchicago.edu

Abstract. Does affective polarization respond symmetrically to a major socio-political shock? Existing work treats partisan animus as a bilateral gap and measures it with periodic cross-sectional surveys, leaving the event-level dynamics and multidimensional structure of polarization underexplored. This study exploits the killing of George Floyd (May 25, 2020) and the subsequent Black Lives Matter (BLM) uprising as a natural quasi-experiment. Using an original corpus of 139,064 tweets collected within a ±100-day window, I infer partisan stance with a zero-shot BART-MNLI classifier (accuracy = 83.7%, macro-F1 = 0.755) and construct four text-based measures of partisan affect (ideological stance, sentiment, moral-foundations language, and destructive toxicity) for both political elites (U.S. Congress, Governors, Presidents) and the mass public. An Interrupted Time Series design, complemented by LOWESS and Gaussian-kernel local linear regression, reveals pronounced asymmetries. Democrats responded through moral intensification and improving sentiment, whereas Republicans responded through sharp ideological consolidation and a gradual escalation of toxicity. Elites amplified the initial framing moment through abrupt level shifts, while the public sustained the trajectory through continuing trends. A moral-foundations decomposition shows that the care/harm domain dominated BLM discourse and persisted longer in Democratic language. Because moralization and toxicity are nearly orthogonal in the corpus ($r \approx -0.08$), they behave as genuinely different affective channels rather than as two readings of one reaction, a structure that aggregate gap scores cannot see.

Keywords: Affective polarization · Social media · Moral foundations · Interrupted time series · Computational social science.

1 Introduction

On May 25, 2020, bystander video recorded Minneapolis police officer Derek Chauvin kneeling on George Floyd's neck for over nine minutes. Within days, protests had spread to all fifty U.S. states; within weeks, "Black Lives Matter" had become the dominant political phrase in American public discourse. Yet on social media the political response split into two starkly different conversations. Democratic politicians and their supporters flooded Twitter with appeals

to justice, harm, and moral duty. Republican politicians and their supporters responded with appeals to order, institutional loyalty, and, over time, escalating hostility. A single event thus generated not a uniform rise in mutual dislike but a divergence whose shape, timing, and rhetorical register differed systematically across the two coalitions.

This paper asks whether affective polarization responds symmetrically to a major socio-political shock. The prevailing literature treats affective polarization as a bilateral phenomenon, a widening gap in which both sides dislike each other in roughly equal measure [23]. When asymmetry receives attention, it is typically confined to the observation that Republicans score slightly higher on out-party hostility, or to debates about whether "both sides" are equally responsible for democratic erosion [29]. This framing obscures a more consequential possibility: that the form, timing, and mechanism of partisan animus may differ systematically between the two parties, and that these differences become especially visible after morally charged events.

Three features of the existing literature leave this possibility unexamined. The standard measurement tools, namely feeling thermometers and social-distance items in periodic surveys, capture the level of partisan dislike at a point in time but cannot observe how partisan affect unfolds around specific events [11]. Most studies treat affective polarization as unidimensional, collapsing moralization, aversion, and othering into a single index despite theoretical arguments that these are empirically separable components with distinct antecedents and consequences [13, 7]. And the relationship between elite and mass polarization has rarely been examined at the temporal resolution that event-level analysis requires [12]. Social-media data, recorded daily and in volume, can address all three gaps, but only when analyzed with temporal designs rather than cross-sectional snapshots.

This study exploits the George Floyd killing as a natural quasi-experiment. Using 139,064 tweets collected within a ± 100-day window around May 25, 2020, I construct four text-based measures of partisan affect for political elites and the mass public: ideological stance, sentiment, moral-foundations language, and destructive toxicity. The identification strategy is an Interrupted Time Series (ITS) design that estimates both immediate level shifts and post-event slope changes for each measure–source–party combination, with LOWESS and kernel local linear regression as robustness checks. A supplementary analysis decomposes moral content using the extended Moral Foundations Dictionary, with attention to the care/harm domain that organizes much of the BLM response.

The response that emerges is asymmetric across affective dimensions. Democrats moralized, raising their moral language and improving their sentiment, whereas Republicans consolidated ideologically and, among the public, escalated toxicity, and the elite shifts were sharper and more abrupt than the sustained public trends. The care/harm foundation that dominated BLM discourse persisted longest in Democratic language. The study thus contributes event-level evidence that partisan animus travels through different affective channels across the two parties, a distinction that aggregate gap scores average away, and, by track-

ing elite and public discourse at daily resolution, it locates the moral register through which BLM charged Democratic identity. The paper proceeds as follows. Section 2 develops the framework and hypotheses; Section 3 describes data and methods; Section 4 presents results; Section 5 discusses implications and limitations.

2 Literature Review and Theoretical Framework

2.1 Affective Polarization, Moralization, and Partisan Asymmetry

Rooted in the growing salience of partisan alignment among co-existing social identities, affective polarization captures positive or negative evaluations of self and others based on partisan membership, regardless of policy substance [21, 22]. As racial, religious, cultural, and ideological identities have stacked onto party labels, inter-party conflict increasingly feels like a contest between ways of life rather than a negotiation over policy [28, 29], and at its extreme it licenses dehumanization and support for political violence [24].

Yet affective polarization has not always been symmetric. Under the dominant framework, which operationalizes polarization as the difference between in-party warmth and out-party coldness, much asymmetry is invisible [11, 12]. The political-sectarianism framework offers a more differentiated view, decomposing partisan animus into othering, aversion, and moralization, each with distinct psychological antecedents and democratic consequences [13]. Recent measurement work validates this decomposition, showing that the subdimensions correlate differently with anti-democratic attitudes and support for political violence [7]. Once affective polarization is treated as multidimensional, asymmetry can emerge in several ways, whether because one side moves faster, moralizes more intensely, or responds more strongly to an identical shock, none of which need appear in an aggregate gap score.

Two families of mechanism explain how such asymmetries develop. One concerns *moralization*, the process by which an attitude comes to be experienced as a matter of right and wrong rather than opinion [34]. Because moralized attitudes resist deliberative revision and reduce tolerance for disagreement, they predict both stronger partisan hostility and reduced willingness to accept adverse election outcomes [33]. Moralization is not uniform, however, since it works as a value-protective response to perceived dyadic harm and intensifies when people see identifiable victims suffering at the hands of identifiable agents [8]. That harm-specific pathway matters here because liberals and conservatives organize moral judgment around different foundations, with liberals weighting care and fairness more heavily while conservatives draw more evenly across loyalty, authority, and sanctity [15], so that when an event centers visible harm to a racial minority by state agents, the liberal framework identifies a victim faster and with greater certainty. Indeed, Democrats' hostility toward Republicans is driven substantially by the perception that Republicans enable harm to disadvantaged groups [30], a perception that events like BLM make vivid and public.

A second concerns *value protection and threat sensitivity.* When the dominant framing instead centers harm to minorities, conservative value hierarchies do not offer the same moral resonance; instead the event threatens values Republicans hold more centrally, among them law enforcement, institutional authority, and public order. The Intergroup Value Protection Model predicts that groups facing perceived threats to core values respond with defensive consolidation, closing ranks and sharpening boundaries [37]. Because conservatives also tend toward stronger needs for cognitive closure, the same identity-mobilizing event can push the two coalitions in structurally different directions.

These dynamics also unfold unevenly over time. Feedback among elites is itself asymmetric, in that Republicans appear to have crossed a threshold where self-reinforcing processes accelerate polarization while Democrats have not [26], and misperception sustains the spiral, as partisans overestimate the opposing party's extremity [1, 25] and may be swayed even more by ingroup norms than by out-group meta-perception [40]. The asymmetries that persist do so by context and content rather than on any single universal pattern, which is why an event-level case can show one coalition moving where aggregate scores would show stalemate. The present study asks whether BLM produces such an asymmetry in affective rather than ideological terms.

2.2 Online Media and the Structure of Asymmetric Escalation

If affective polarization is dynamic rather than fixed, the environment in which responses take shape matters, and social media is where elite frames, public reactions, moral appeals, and hostility are produced and recirculated. Its effect on polarization is real but conditional, with platforms acting as accelerants for existing moral dynamics rather than independent causes [36]. Two features of that environment bear directly on event-driven asymmetry. One is partisan sorting: digital media divide partisans not by sealing them in ideological bubbles but by aligning their social, cultural, and political differences along party lines [35], and the sorting is itself lopsided. Ideological segregation on Facebook during 2020 featured a homogeneously conservative enclave with no liberal equivalent [14], and field experiments that exposed partisans to opposing-party messaging moved Republicans further right without comparably shifting Democrats [2], so identical exposure shocks can yield different affective outcomes across coalitions.

The other is that the platform rewards morally and emotionally charged content, since moral-emotional language travels furthest within ideological networks [4] and out-group animosity is the single strongest predictor of engagement [32]. These incentives meet an asymmetric elite ecology in which Republican media engagement in Congress grew more homophilous and extreme through the 2010s [17]. Although most users do not follow elites directly [39], elite frames travel through journalists, activists, and ordinary users, so that the emotional tone elites strike is echoed in public affect. Because exposure to hostile partisan messaging widens perceived ideological distance the longer it is processed [3] and partisans overperceive the outrage around them [5], observed outrage breeds

perceived outrage and then further real outrage, a feedback loop anchored in the reciprocal link between perceived polarization and moralization [9].

2.3 The BLM Case and Hypotheses

The George Floyd killing makes these mechanisms visible. BLM discourse was organized around visible bodily harm, racial vulnerability, police power, and whether institutions had failed in their duty to protect, all of which are moral elements that map directly onto care and fairness, so that the event was pre-encoded in the language the Democratic coalition finds most compelling. Care/harm was the dominant moral foundation in post-Floyd Twitter discourse and moral language tracked pro-BLM framing [31], while partisan identity was the strongest predictor of whether citizens read the protests as legitimate grievance or as a threat to public order [10]. The case is valuable not because it is the only event that could produce asymmetry but because its particular fusion of moral content and partisan identity makes the moralization and value-protection mechanisms unusually legible, yielding three testable expectations.

H1 (Asymmetric affective response). The shock raises moral-foundations language more among Democrats, reflecting the greater salience of racialized harm in their value hierarchy, while Republicans respond through ideological consolidation and gradual toxicity escalation, reflecting value-protective defense of authority and institutional loyalty under threat.

H2 (Elite amplification). Elite affective shifts exceed the corresponding public shifts in magnitude, reflecting the pressure on officeholders to define the event and supply rhetoric that later diffuses through online publics [41, 17].

H3 (Temporal divergence). Elite responses take the form of abrupt level shifts without sustained trend change, whereas public responses combine an initial shock with a continuing trend, consistent with layered diffusion and the self-reinforcing dynamics of moralization and perceived polarization [9, 5].

3 Data and Methods

3.1 Data Sources

The corpus spans a ±100-day window around May 25, 2020 (February 14–September 2), capturing both the pre-event baseline and the post-event reaction. Over 160,000 tweets were retrieved through a third-party Twitter API using BLM-related keyword queries (e.g., "Black Lives Matter," "George Floyd," "police brutality," "defund the police," "systemic racism") with noise-exclusion terms and a retweet filter. Elite tweets came from the official accounts of sitting members of the U.S. Congress, state Governors, and Presidents of both parties; after deduplication, English-only filtering, and removal of tweets referencing other victims (Breonna Taylor, Ahmaud Arbery) to isolate the Floyd shock, the elite corpus holds 1,672 tweets (1,384 Democratic; 288 Republican), whose verified party affiliations serve as ground truth for the classifier. Public

tweets, sampled at up to 800 per day and subjected to the same filtering plus bot removal (accounts younger than one day), number 137,392; lacking verified labels, their partisan stance is inferred by zero-shot classification (Section 3.3). The combined analytic dataset totals $N = 139{,}064$.

One structural feature shapes the analysis. Because BLM keywords were rare in political discourse before Floyd's death, the elite pre-event sample is thin, fewer than fifteen tweets per party (Table 3), so the elite pre-trends rest on little data, whereas the public baseline of over 69,000 pre-event tweets is stable. As Section 4 shows, that sparsity is not merely a limitation but part of the finding: there was scarcely any elite conversation for the event to interrupt.

3.2 Measures of Affective Polarization

Four text-based measures, each computed per tweet and aggregated to daily means, capture distinct dimensions of partisan affect. *Ideological stance* is a continuous score $P(\text{Dem})$ from a zero-shot BART-MNLI classifier [27] contrasting a Democratic frame ("blaming systemic racism and police brutality for the tragedy") with a Republican one ("blaming riots, looters, or radical anti-police demands"); tweets at or above 0.5 are labeled Democratic. *Sentiment* is the VADER compound score [19], from -1 to $+1$, designed for social-media text. *Moral-foundations language* comes from the extended Moral Foundations Dictionary [18], which scores each tweet on five foundations (care, fairness, loyalty, authority, sanctity) along probability and sentiment dimensions; the aggregate index combines these into a continuous measure of moral intensity, and the foundation-specific probabilities and a targeted harm-framing index feed the supplementary moral analysis. *Destructive toxicity* is the Detoxify Toxic-BERT probability [16], on a 0–1 scale, where high values mark severely toxic, threatening, or identity-attacking language.

3.3 Stance Classification and Validation

Partisan stance is inferred using zero-shot classification with BART-MNLI. Multiple prompt strategies were evaluated on the elite ground-truth sample, including framing-theory, moral-foundation, and narrative-focus prompts; the framing-theory prompt achieved the strongest balanced performance across parties. Table 1 reports performance on the elite ground-truth sample ($N = 1{,}672$). Overall accuracy is 83.7%, with macro-F1 $= 0.755$ and ROC AUC $= 0.880$. Democratic tweets achieve F1 $= 0.897$ (precision $= 0.942$, recall $= 0.856$), while Republican tweets achieve F1 $= 0.613$ (precision $= 0.519$, recall $= 0.747$). The lower Republican recall, which arises because bipartisan language around policing reform pulls some Republican tweets toward the Democratic frame, means the classifier may underrepresent moderate Republican voices, though performance stays stable across the post-event window (weekly accuracy 77.5–93.2%).

Table 1. Stance classifier validation on elite ground truth.

Class / Statistic	Precision	Recall	F1	Support
Democratic	0.942	0.856	0.897	1,384
Republican	0.519	0.747	0.613	288
Macro average	0.731	0.801	0.755	1,672
Weighted average	0.869	0.837	0.848	1,672
Accuracy			0.837	1,672
ROC AUC			0.880	
Tuned threshold			0.528	

3.4 Identification Strategy

The core strategy is an Interrupted Time Series design estimated via ordinary least squares, complemented by two non-parametric approaches as robustness checks. Models are estimated separately for each combination of measure, source, and party, yielding 16 core specifications per estimation approach.

OLS Interrupted Time Series. The primary specification models the daily aggregated outcome as a piecewise-linear function of time with a structural break at the intervention date:

$$Y_t = \beta_0 + \beta_1 \,\mathrm{Time}_t + \beta_2 \,\mathrm{Intervention}_t + \beta_3 \,\mathrm{TimeAfter}_t + \varepsilon_t, \tag{1}$$

where Y_t is the daily mean of the affective measure for a given party–source cell, Time_t the day index since the window's start, $\mathrm{Intervention}_t$ an indicator equal to 1 on or after May 25, and $\mathrm{TimeAfter}_t$ the count of post-event days (0 before). The coefficient β_2 captures the immediate shock and β_3 the change in trend slope. Two non-parametric checks relax the linearity assumption: LOWESS smoothing fit separately on each side of the cutoff (bandwidth $f = 0.18$), with local linear slopes extracted in ± 14-day windows, and Gaussian-kernel local linear regression in the regression-discontinuity tradition [20], with bandwidth $h = 30$ days and weighted least squares with HC1 robust errors, so that observations nearest the cutoff receive the greatest weight. The supplementary moral analysis adds a 7-day rolling mean of harm framing, a feature-profile heatmap of group means and within-feature z-scores, and a Pearson correlation matrix of the text measures.

4 Results

Table 2 reports the OLS ITS estimates, and the argument of this section builds from them in three steps: first the central asymmetry between a Democratic moral channel and a Republican consolidation channel, then the content and the temporal form that explain why the split fell as it did, and finally the evidence that these channels are genuinely distinct. Significant effects concentrate in the public series, where the daily samples are dense; the elite estimates, resting on a handful of pre-event tweets, enter mainly as sharp framing shocks.

4.1 Two Baselines Before the Shock

Two structural facts hold before May 25 and frame everything that follows. Ideology separates cleanly by party in both sources, validating the classifier on its face, and elites speak far less toxically than the public (median toxicity ≈ 0.003 vs. ≈ 0.02), an institutional-norm gap that will govern where each group's response can surface. The pre/post means already preview the asymmetry: moral-foundations language rose for Democrats and fell for Republicans in both sources (public $+0.145$ vs. -0.119), and elite sentiment swung nearly a full point apart, Democrats improving while Republicans collapsed. The model-based estimates make these raw contrasts precise.

4.2 The Central Asymmetry: A Moral Channel and a Consolidation Channel

The organizing finding in Table 2 is that the two coalitions engaged the moral register in opposite directions, and that the surplus energy of each side's response then surfaced on a different channel. Among Democrats the response was built around moral affirmation. Elite Democrats jumped at the break in moral-foundations language ($\beta_2 = 0.372$, $p = .006$) with no subsequent trend, in a one-time mobilization of officeholders (Figure 1), while public Democrats combined an immediate rise ($\beta_2 = 0.146$, $p < .001$) with a slope that kept climbing for the remainder of the window ($\beta_3 = 0.0016$, $p < .001$; Figure 1). Sentiment moved with it, as public Democratic valence rose at the break and continued upward, the signature of outrage voiced as righteous solidarity rather than diffuse negativity. Where the Democratic response intensified moral-affirmative language, it did not escalate hostility; public Democratic toxicity, if anything, fell ($\beta_2 = -0.031$, $p = .002$).

For Republicans the moral register moved the other way, because public Republican moral-foundations language dropped at the break ($\beta_2 = -0.072$, $p = .002$) before a partial recovery, and the force of their response surfaced instead on the two channels the Democratic reaction left untouched. Ideological language consolidated, as public Republican $P(\text{Dem})$ fell sharply and kept falling ($\beta_2 = -0.023$, $\beta_3 = -0.0007$, both $p < .001$), and the piecewise-linear model absorbed an unusually large share of the daily variance ($R^2 = 0.78$), a degree of collective regularity rare in social-media series. Toxicity, flat at the break, then climbed gradually over the following weeks ($\beta_3 = 0.0002$, $p = .042$). Sharpest of all is elite Republican sentiment, which collapsed from a positive pre-event mean to negative ($\beta_2 = -0.546$, the largest coefficient in the study) even as elite toxicity held flat, so that officeholders registered alarm through valence rather than destructive speech. What separates the coalitions is therefore kind rather than degree, since one intensified moral-affirmative language while the other tightened partisan-ideological boundaries and, more slowly, hardened in tone. And because elite ideology and all elite toxicity effects were null, the asymmetry runs specifically along the moral and consolidation channels rather than across measures uniformly, which is the pattern H1 anticipated.

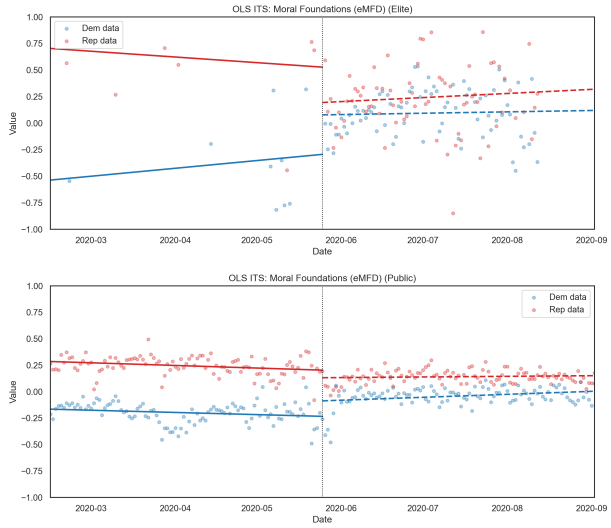

Fig. 1. Interrupted time series of moral-foundations (eMFD) language among elites (top) and the public (bottom). Democratic elites jump sharply at the intervention while the Republican elite effect is negative and non-significant; among the public, Democrats show both an immediate break and a sustained upward trend whereas Republicans show an initial negative break with partial recovery.

Table 2. Interrupted time-series estimates. Standard errors in parentheses. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Measure	Source	Party	Break β_2 (SE)	Post slope β_3 (SE)	N	R^2
Moral Found. (eMFD)	Elite	Dem	0.372^{**} (.135)	-0.0020 (.0036)	85	0.244
Moral Found. (eMFD)	Elite	Rep	-0.333 (.244)	0.0030 (.0047)	81	0.085
Moral Found. (eMFD)	Public	Dem	0.146^{***} (.026)	0.0016^{***} (.0005)	201	0.441
Moral Found. (eMFD)	Public	Rep	-0.072^{***} (.021)	0.0010^{**} (.0004)	201	0.350
Sentiment (VADER)	Elite	Dem	0.216 (.185)	-0.0021 (.0049)	85	0.071
Sentiment (VADER)	Elite	Rep	-0.546 (.284)	0.0017 (.0054)	81	0.091
Sentiment (VADER)	Public	Dem	0.142^{***} (.020)	0.0008^{*} (.0003)	201	0.567
Sentiment (VADER)	Public	Rep	-0.024 (.016)	-0.0001 (.0003)	201	0.364
Toxicity (BERT)	Elite	Dem	-0.027 (.021)	-0.0007 (.0006)	85	0.116
Toxicity (BERT)	Elite	Rep	-0.014 (.046)	0.0005 (.0009)	81	0.076
Toxicity (BERT)	Public	Dem	-0.031^{**} (.010)	-0.0001 (.0002)	201	0.153
Toxicity (BERT)	Public	Rep	-0.003 (.007)	0.0002^{*} (.0001)	201	0.351
Ideology (P(Dem))	Elite	Dem	0.036 (.034)	-0.0017 (.0009)	85	0.137
Ideology (P(Dem))	Elite	Rep	-0.052 (.064)	-0.0015 (.0012)	81	0.084
Ideology (P(Dem))	Public	Dem	-0.021^{***} (.005)	-0.0004^{***} (.0001)	201	0.492
Ideology (P(Dem))	Public	Rep	-0.023^{***} (.005)	-0.0007^{***} (.0001)	201	0.782

4.3 Who Moves First, and Who Sustains It

Elite and public series differ not only in magnitude but in temporal shape, and that difference clarifies the division of labor between the two tiers. Elite effects are

exclusively shock-like, a sharp break with no significant post-event slope, whereas the public effects that reach significance, namely Democratic moral language, sentiment, and ideology, pair an immediate break with a continuing trend. Since the two biggest elite coefficients (the Democratic surge in moral language, the Republican collapse in sentiment) also exceed their public counterparts, elites supply the loudest opening signal while the public carries the trajectory forward. That is a narrower reading of elite amplification (H2) than a claim that elite effects are uniformly larger, since amplification works on the framing moment rather than the whole arc. Even the shock form is diagnostic rather than a nuisance, because with fewer than fifteen elite tweets per party before May 25 (Table 3) there was no standing elite conversation for the event to accelerate, so that BLM created one, and a process with no pre-trend can only register as a break.

Table 3. Source support diagnostics for event-time estimates.

Source	Party	Tweets pre	Tweets post	Daily pre	Daily post	14d pre	14d post
Elite	Dem	11	1,208	9	76	3	14
Elite	Rep	9	444	8	73	3	14
Public	Dem	31,624	25,707	100	101	15	15
Public	Rep	37,618	42,443	100	101	15	15

4.4 Why the Moral Channel Belonged to Democrats

If the Democratic response ran through moralization, the content of that moralization should reveal why this channel was available to Democrats and not Republicans. Decomposing the moral language confirms that BLM activated one foundation in particular, since across all groups care/harm was the most prominent (0.122), ahead of authority (0.106), fairness (0.105), loyalty (0.105), and sanctity (0.093). The event spoke in the care idiom, and Democrats spoke it more fluently, scoring higher than Republicans on every foundation but most widely on the foundations that organize the liberal moral palette, care (public Dem 0.131 vs. Rep 0.114) and fairness (0.112 vs. 0.099). Among elites the party gap on care nearly vanished (0.125 vs. 0.124), as if officeholders of both parties were pulled onto care terrain by the event itself while the public split along the predicted partisan lines.

A targeted harm-framing index, defined as care probability weighted by negative care sentiment ($\text{harm} = \text{care}_p \times \max(-\text{care}_{\text{sent}}, 0)$), sharpens the dynamic. Public Democratic harm framing ran above the Republican level throughout the window, spiked at May 25, and stayed elevated for the rest of the study, whereas Republican harm framing spiked only briefly before returning to baseline within about three weeks, and the elite series showed the same contrast in compressed

form. This persistence mirrors the aggregate eMFD result, in which the sustained Democratic slope ($\beta_3 = 0.0016$, $p < .001$) stands against the Republican shock-and-recovery, so that a measure built specifically on harm and one built on moral language as a whole agree that BLM discourse was organized around care and engaged more intensely and durably by Democrats.

4.5 Are the Channels Genuinely Distinct?

The two-channel reading rests on one measurement fact: moralization and toxicity must not be redundant, or the Democratic and Republican patterns would collapse into two readings of a single underlying hostility. They do not. Across the corpus, toxicity is nearly orthogonal to aggregate moral-foundations language ($r = -0.08$) and only weakly tied to sentiment ($r = -0.23$), whereas moral language and sentiment share substantial variance ($r = 0.57$). A tweet could therefore be intensely moralized without being toxic, or toxic without invoking moral foundations at all. This near-independence is what licenses the central claim: the Democratic moral channel and the Republican toxicity channel are not the same affect measured twice but distinct registers that the event activated for different coalitions.

4.6 Robustness

The public-sample findings survive both non-parametric checks, so they are not artifacts of the OLS linearity assumption. LOWESS reproduces the significant post-event slope changes in public Democratic ideology, sentiment, and moral language, and adds one telling detail, that public Democratic toxicity carried a significant rising pre-trend (0.004, $p = .007$) which the event reversed, consistent with BLM interrupting an escalation already under way. The kernel estimator, which weights observations near the cutoff most heavily, confirms the Republican ideology and Democratic moral and sentiment slopes and sharpens the Republican toxicity result (0.0008, $p < .001$ vs. OLS 0.0002, $p = .042$), locating that escalation in the weeks immediately after the event.

5 Discussion

The results pose a question the levels-based tradition in affective-polarization research is not built to answer. If partisan animus were a single quantity that rises and falls, a shared shock should push both coalitions in the same direction and differ at most in degree; BLM instead moved the two sides through different registers, one moral and one ideological-and-hostile, a divergence in kind that aggregate gap scores are designed to average away. The sections below read that divergence against the multidimensional account of partisan animus, the moral psychology of value protection, and the dynamical models of how an asymmetry hardens.

5.1 Two Kinds of Negativity

The statistical independence of the moral and toxic channels (Section 4.5) is what gives the political-sectarianism framework [13, 7] its purchase here. What that literature establishes on conceptual and survey grounds, namely that othering, aversion, and moralization are separable subdimensions of partisan hostility, the BLM episode shows emerging dynamically, so that a measurement claim becomes a claim about mechanism, since a single stimulus can engage the subdimensions selectively and load onto different ones for different coalitions.

Reading the two responses as quantities of the same hostility misses that they are affectively opposite. Democratic moral language carried resolve and solidarity alongside anger, which is the profile of outrage at perceived injustice [4], while Republican elites signaled alarm at a movement they experienced as a threat to institutional authority and public order. This is precisely the reaction van Zomeren's value-protection model [37] anticipates when a group's core values, rather than its policy preferences, are placed under threat, namely defensive consolidation rather than moral counter-mobilization. That the same alarm surfaced as collapsing sentiment among norm-constrained officeholders but as rising toxicity among the unconstrained public locates the institutional boundary exactly where the pre-event baseline placed it, and turns a descriptive gap into part of the explanation.

The sharpest dialogue is with the moralization literature itself. Skitka [34] and Ryan [33] have established moralization as a source of rigidity, intolerance, and refusal to compromise, and the present case disputes none of those downstream costs. But the Democratic trajectory, in which moral language rose as toxicity fell and the pre-event toxic trend uncovered by the robustness analysis reversed, points to a function the literature has not emphasized, namely that moralization can discipline the *form* of partisan expression. A clear narrative built around an identifiable victim and an identifiable agent of harm [8] gives outrage a channel, converting what might otherwise be diffuse aggression into principled claim-making. On this reading moralization can make discourse more intense and less crude at once, a possibility that matters because it breaks the usual assumption that moral and hostile escalation move together.

5.2 Why the Asymmetry Hardened

An initial divergence need not last; the more revealing fact is that this one deepened along the very channel each coalition had opened. The Republican sequence of boundary closure before hostility follows the logic of threat response, in which groups consolidate identity before escalating outward, and aligns with evidence that exposure to hostile partisan messaging widens perceived ideological distance the longer it is processed [3]. Persistence itself invites a feedback account, since the moralization–polarization loop [9] and the systematic overperception of online outrage [5] can each sustain an escalating spiral. A general amplifier, however, would raise every channel together, and the Democratic series shows the opposite, with moral language and ideology climbing while toxicity held flat,

so that the feedback must be channel-selective. Through the nonlinear-dynamics model of Leonard et al. [26], the event fixes the system's initial condition on one axis and the self-reinforcing dynamics then run along that axis rather than across all of them. Which form of polarization endures is therefore settled at the moment of activation, by the fit between event content and the value hierarchy it engages, not by a standing disposition of one party to be "more polarized" than the other.

5.3 Implications

Three debates look different in this light. For measurement, the finding that one shock moves the subdimensions of sectarianism in opposite directions makes multidimensional measurement a design requirement for event-level work, not a robustness afterthought [7]. For intervention, the independence of the moral and toxic channels implies that reducing moralization and reducing toxicity are distinct tasks [38], because misperception correction [1, 25] can address hostility built on exaggerated beliefs but bites less on moralization grounded in the partly accurate perception that the other side enables harm [30], and because cross-cutting exposure moderates in some settings yet hardens in others [6, 2], so that diagnosis should be channel-specific. For the sites of study, protest waves emerge as natural experiments on affect, since they are compressed, highly visible, and force each party to declare which victims deserve sympathy [10]. Because the direction of the asymmetry tracks the moral content of the trigger, an event built on threats to authority or sanctity should invert it and produce Republican moralization against Democratic defensive consolidation, which is a directional prediction the framework can be tested against.

5.4 Scope and Limits

These claims are bounded in two ways. Because Floyd's killing was built from care/harm content, the direction of the asymmetry follows from that content, so an event organized around loyalty or sanctity could invert it, a prediction the framework generates rather than a weakness it hides. The window also overlaps COVID-19 and the 2020 campaign, which leaves the extended post-event slopes, though not the sharp breaks at the cutoff, exposed to concurrent shocks. A second limit lies in the medium. Because the classifier reads Democratic discourse more reliably than Republican (F1 0.897 vs. 0.613), it likely flattens internal Republican variation, and Twitter's users skew young, engaged, and extreme [39]. What the text records, finally, is the rhetorical signature of moralization rather than private conviction, a psychology that should travel across platforms even where the amplification dynamics shaping this corpus do not.

Within those bounds, the contribution is to show that affective polarization, observed at the event level with multidimensional measures, runs along separable channels that aggregate indices average out, and that the fit between event content and partisan value hierarchies decides both which asymmetry emerges and which feedback sustains it.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Ahler, D.J., Sood, G.: The parties in our heads: misperceptions about party composition and their consequences. J. Polit. **80**(3), 964–981 (2018)
2. Bail, C.A., et al.: Exposure to opposing views on social media can increase political polarization. Proc. Natl. Acad. Sci. **115**(37), 9216–9221 (2018)
3. Banks, A., Calvo, E., Karol, D., Telhami, S.: #PolarizedFeeds: three experiments on polarization, framing, and social media. Int. J. Press/Politics **26**(3), 609–634 (2021)
4. Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J.: Emotion shapes the diffusion of moralized content in social networks. Proc. Natl. Acad. Sci. **114**(28), 7313–7318 (2017)
5. Brady, W.J., McLoughlin, K., Torres, M.P., Luo, K.F., Gao, M., Crockett, M.J.: Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. Nat. Hum. Behav. **7**(6), 917–927 (2023)
6. Broockman, D.E., Kalla, J.L.: Consuming cross-cutting media causes learning and moderates attitudes: a field experiment with Fox News viewers. J. Polit. **87**(1), 246–261 (2025)
7. Campos, N., Federico, C.: A new measure of affective polarization. Am. Polit. Sci. Rev., 1–19 (2025)
8. D'Amore, C., van Zomeren, M., Koudenburg, N.: How moralization of dyadic harm leads to action against, and for, the other. Eur. J. Soc. Psychol. **52**(5–6), 896–910 (2022)
9. D'Amore, C., van Zomeren, M., Koudenburg, N.: Affective polarization and attitude moralization: a four-wave longitudinal study during the 2020 U.S. election. Group Process. Intergroup Relat. (2024)
10. Drakulich, K.M., Denver, M.: Views of the George Floyd and Black Lives Matter protests: racial and partisan differences. Sociol. Race Ethn. **8**(3), 356–372 (2022)
11. Druckman, J.N., Levendusky, M.S.: What do we measure when we measure affective polarization? Public Opin. Q. **83**(1), 114–122 (2019)
12. Druckman, J.N., Klar, S., Krupnikov, Y., Levendusky, M., Ryan, J.B.: (Mis)estimating affective polarization. J. Polit. **83**(3), 1176–1180 (2021)
13. Finkel, E.J., et al.: Political sectarianism in America. Science **370**(6516), 533–536 (2020)
14. González-Bailón, S., et al.: Asymmetric ideological segregation in exposure to political news on Facebook. Science **381**(6656), 392–398 (2023)
15. Graham, J., Haidt, J., Nosek, B.A.: Liberals and conservatives rely on different sets of moral foundations. J. Pers. Soc. Psychol. **96**(5), 1029–1046 (2009)
16. Hanu, L., Unitary team: Detoxify. GitHub (2020), `https://github.com/unitaryai/detoxify`
17. Heseltine, M.: Asymmetric polarization in online media engagement in the United States Congress. Int. J. Press/Politics **30**(1), 299–325 (2025)
18. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended Moral Foundations Dictionary (eMFD). Behav. Res. Methods **53**, 232–246 (2021)
19. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proc. Int. AAAI Conf. Web Soc. Media, vol. 8, pp. 216–225 (2014)

20. Imbens, G.W., Lemieux, T.: Regression discontinuity designs: a guide to practice. J. Econom. **142**(2), 615–635 (2008)
21. Iyengar, S., Sood, G., Lelkes, Y.: Affect, not ideology: a social identity perspective on polarization. Public Opin. Q. **76**(3), 405–431 (2012)
22. Iyengar, S., Westwood, S.J.: Fear and loathing across party lines: new evidence on group polarization. Am. J. Polit. Sci. **59**(3), 690–707 (2015)
23. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., Westwood, S.J.: The origins and consequences of affective polarization in the United States. Annu. Rev. Polit. Sci. **22**(1), 129–146 (2019)
24. Kalmoe, N.P., Mason, L.: Radical American Partisanship. University of Chicago Press, Chicago (2022)
25. Lees, J., Cikara, M.: Understanding and combating misperceived polarization. Philos. Trans. R. Soc. B **376**(1822), 20200143 (2021)
26. Leonard, N.E., Lipsitz, K., Bizyaeva, A., Franci, A., Lelkes, Y.: The nonlinear feedback dynamics of asymmetric political polarization. Proc. Natl. Acad. Sci. **118**(50), e2102149118 (2021)
27. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proc. 58th Annu. Meet. ACL, pp. 7871–7880 (2020)
28. Mason, L.: Ideologues without issues: the polarizing consequences of ideological identities. Public Opin. Q. **82**(S1), 866–887 (2018)
29. Mason, L.: Uncivil Agreement: How Politics Became Our Identity. University of Chicago Press, Chicago (2018)
30. Nair, K., Anderson, R.A., Spelman, T., Mosleh, M., Kouchaki, M.: Asymmetric polarization: the perception that Republicans pose harm to disadvantaged groups drives Democrats' greater dislike of Republicans. J. Pers. Soc. Psychol. (2025)
31. Priniski, J.H., Horne, Z., Yang, K.: The moral foundations of the Black Lives Matter movement on Twitter. In: Proc. Annu. Conf. Cogn. Sci. Soc. (2021)
32. Rathje, S., Van Bavel, J.J., Van der Linden, S.: Out-group animosity drives engagement on social media. Proc. Natl. Acad. Sci. **118**(26), e2024292118 (2021)
33. Ryan, T.J.: No compromise: political consequences of moralized attitudes. Am. J. Polit. Sci. **61**(2), 409–423 (2017)
34. Skitka, L.J.: The psychology of moral conviction. Soc. Personal. Psychol. Compass **4**(4), 267–281 (2010)
35. Törnberg, P.: How digital media drive affective polarization through partisan sorting. Proc. Natl. Acad. Sci. **119**(42), e2207159119 (2022)
36. Van Bavel, J.J., Robertson, C.E., Del Rosario, K., Rasmussen, J., Rathje, S.: Social media and morality. Annu. Rev. Psychol. **75**, 311–340 (2024)
37. van Zomeren, M.: The Intergroup Value Protection Model. Eur. Rev. Soc. Psychol. (2024)
38. Voelkel, J.G., et al.: Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes. Nat. Hum. Behav. **7**(1), 55–64 (2022)
39. Wojcieszak, M., Casas, A., Yu, X., Nagler, J., Tucker, J.A.: Most users do not follow political elites on Twitter. Sci. Adv. **8**(32), eabn9418 (2022)
40. You, Z.T., Lee, S.W.S.: Explanations of and interventions against affective polarization cannot afford to ignore the power of ingroup norm perception. PNAS Nexus **3**(10), pgae286 (2024)
41. Zaller, J.R.: The Nature and Origins of Mass Opinion. Cambridge University Press, Cambridge (1992)