Northwestern | INSTITUTE FOR POLICY RESEARCH

RESEARCH EXCELLENCE • POLICY IMPACT

IPR Working Paper Series

WP-25-31

Why Reform Stalls: Justification and Outrage as Competing Public Responses to Police Violence

Jonathan Doriscar
Northwestern University and IPR

Ava Ma de Sousa
University of California, Santa Barbara

Lauryn Hoard
The George Washington University

Wendi Gardner
Northwestern University

William Brady
Northwestern University and IPR

Sylvia Perry
Northwestern University and IPR

Version: December 23, 2025

DRAFT

Please do not quote or distribute without permission.

Abstract

Across two studies, the researchers investigate how public justifications of police violence and moral outrage emerge as competing responses with distinct implications for reform. In Study 1, they analyzed 257,401 comments from 57 widely viewed YouTube videos using large language models to detect justificatory rhetoric, expressions of outrage, and calls for accountability. Justifications were consistently associated with reduced reform-oriented discourse, whereas outrage strongly co-occurred with accountability language. When modeled together, each response showed stronger links to reform, consistent with mutual suppression and the idea that they function as opposing interpretive strategies. In Study 2, the researchers experimentally manipulated victim race by exposing participants (N = 159) to severity-matched videos of police brutality. Justification and outrage again diverged, with justification associated with lower reform support and outrage associated with higher reform support. Participants were more likely to attribute superhuman traits to the Black victim—a racialized justification tied to reduced reform support and to weaker translation of outrage into reform. Outrage for the Black victim was also a more consistent predictor of reform than outrage for the White victim. Together, these findings show how justification and moral outrage function as competing responses to police violence, and how racialized justifications can shape the conditions under which outrage mobilizes collective demands for change.

Acknowledgements and Disclosures

The authors gratefully acknowledge the Social Cognition and Intergroup Processes Lab at Northwestern University for supporting this research, particularly the many undergraduate research assistants who contributed countless hours to video coding. They are especially thankful to research assistants Yadira Lopez, Shaniya Kendrick, Narmeen Charhal, and Jacob Hodges, whose dedication helped bring this project to life. We also thank the Social Self Lab (Northwestern Psychology) and the Social Learning and Algorithmic Bias Lab (Kellogg School of Management) for their constructive feedback on presentations and early manuscript drafts. Jonathan Doriscar would also like to thank the Black Graduate Conference in Psychology for offering a formative space to grow this work in dialogue with like-minded scholars and mentors. This research was supported by the National Science Foundation (NSF) Graduate Research Fellowship awarded to Jonathan Doriscar and by the Institute for Policy Research at Northwestern University. Jonathan Doriscar extends heartfelt thanks to his family—Emmanuelle, Jude, Elijah, Danielle, Christine, Amanda, and Karina—for being a constant source of encouragement and support throughout this project. The authors also thank YouTube for providing access to their API, which made the large-scale data collection for this project possible. Finally, they acknowledge the countless victims of police brutality whose lives and stories form the foundation of this research.

All data files and code are provided openly here: https://osf.io/8wczy/

The authors declare no competing interests.

Author Contributions: Jonathan E. Doriscar conceived of and led the project, and contributed to conceptualization, methodology, data curation, formal analysis, validation, investigation, visualization, writing – original draft, and writing – review and editing. Ava Q. Ma de Sousa and Lauryn Hoard contributed to conceptualization, data curation, and writing – review and editing. William J. Brady contributed to conceptualization, data curation, formal analysis, methodology, validation, supervision, and writing – review and editing. Wendi L. Gardner and Sylvia P. Perry contributed to conceptualization, methodology, formal analysis, validation, supervision, and writing – review and editing. William J. Brady contributed to conceptualization, data curation, formal analysis, methodology, validation, supervision, and writing – review and editing.

Why Reform Stalls: Justifications of Force Are Linked to Lower Outrage and Reform Support

Why do some incidents of police violence provoke widespread outrage and reform efforts, while others fade into silence? In the United States, police brutality—the excessive use of force by law enforcement—remains a deeply entrenched and highly visible issue. In 2023 alone, police killed 1,353 people, marking the deadliest year in over a decade[1]. Despite growing public scrutiny, bipartisan reform proposals, and widespread adoption of body cameras, the rate of fatal police encounters has remained alarmingly stable. Yet public support for police reform remains inconsistent. Some incidents trigger collective outrage and national action; others do not. What accounts for this divergence? We consider the possibility that public responses bifurcate into competing rhetorical–emotional modes—justifying officers' force versus expressing moral outrage—which, in turn, shape support for reform (see Figure 1).

One promising explanation lies in the emotional and interpretive processes that follow acts of violence. Moral outrage—a strong emotional response to perceived violations of fairness and justice—plays a critical role in mobilizing support for social change[2, 3]. Outrage often motivates political behaviors such as protesting, donating, and calling for institutional accountability[4,5]. In the context of police brutality, public outrage has played an important role in catalyzing policy shifts and reform movements[6]. Public reactions often diverge: some respond with outrage, others with justification. When observers—whether police, pundits, or members of the public—frame violent encounters as *justified*, the moral force of the event may be neutralized. Justifications of police officers' use of excessive force can shift blame away from law enforcement by reframing victims as threatening aggressors—individuals who posed a

danger to officers and therefore warranted force—and in doing so, offer an alternative frame—one less tied to the emotional urgency that fuels reform[7].

These justification narratives often rely on recurring rhetorical patterns: appeals to officer fear (e.g., "I feared for my life"), emphasis on victim noncompliance, or characterizations of the victim as dangerous or deviant[8,9,10,11]. Such narratives are common. Following the 2016 killing of Daniel Shaver—who was unarmed and sobbing on the floor when police shot him five times—the officer described him as noncompliant and possibly reaching for a weapon[12]. In 2017, Justine Damond was shot by an officer who later explained he had been "startled" and acted in a "split-second decision"[13]. Breonna Taylor's killing was justified post hoc as the outcome of a narcotics investigation, despite the suspect already being in custody and no drugs found at the scene[14]. The murder of George Floyd in 2020 catalyzed one of the largest protest movements in U.S. history, sparking a global reckoning with police violence and racial injustice. Yet even five years later, public narratives attempting to reframe his death continue to circulate. Some commentators and even judicial figures have suggested that Floyd died of a drug overdose or emphasized his criminal record—justifications that shift attention away from police violence and serve to dampen moral outrage[14, 15].

Although justifications can be drawn on by people across demographics and for varied motivations, the form they take often depends on the identity of the victim. In the U.S., Black Americans are three times more likely to be killed by police than White individuals, and 1.3 times more likely to be unarmed when killed[11,16,17]. Black victims are also disproportionately targeted by racialized stereotype content, including superhumanization[18]—the belief that Black individuals possess extraordinary strength, emotional restraint, and resilience, rendering them less vulnerable to harm and more deserving of force[19, 20]. The use of such stereotypes has been

linked to distorted perceptions of Black victims, reduced empathy, and increased justification of violence—especially in criminal justice contexts[21]. A striking example can be found in Officer Darren Wilson's description of Michael Brown: "The only way I can describe it is I felt like a five-year-old holding on to Hulk Hogan"[22]. We therefore do not anticipate a uniform main effect of victim race on justification overall. Instead, we predict that racialized justifications—particularly superhumanization—will be more strongly applied to Black victims and will carry distinct consequences, helping explain when public reactions diverge between justification of excessive force by police and moral outrage, with downstream implications for support for police reform[18,19,21]. To test this possibility, Study 1 treats victim race as exploratory at the video level, given limited between-video power, while Study 2 manipulates victim race to assess whether superhumanization of Black (versus White) victims is disproportionately applied, and how this response stands in contrast to outrage in its implications for reform.

Present Research

Across two studies—one observational and one experimental—we investigated how public justifications of excessive force by police relate to moral outrage and public calls for accountability and police reform. In Study 1, we analyzed 257,401 YouTube comments across 57 widely viewed videos (mean views ≈ 2.6M) depicting excessive force to estimate the prevalence of justification narratives (e.g., appeals to fear, accusations of noncompliance, self-defense frames) and their associations with moral outrage and calls for police accountability.

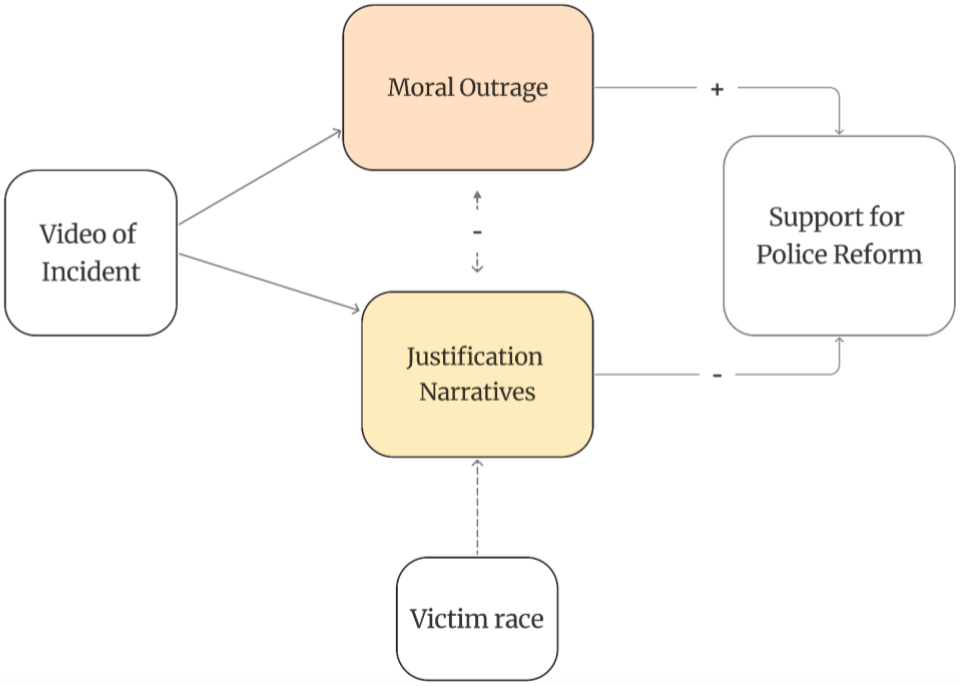
We also explored whether these patterns varied by victim race at the video level, treating race as exploratory given the limited power to detect between-video racial differences. Study 2 built on this framework by manipulating victim race in severity-matched videos and testing three key predictions. First, we expected justification and outrage to emerge as distinct, negatively

related responses—competing ways of interpreting the same events. Second, we expected these responses to diverge in their implications for reform: outrage would be positively associated with support for police reform, while justification would be negatively associated. Third, we expected racialized justifications to be especially likely when the victim was Black. In particular, we examined whether superhumanization—a stereotype portraying Black victims as extraordinarily strong and less vulnerable—would be more strongly endorsed for Black victims and whether this form of justification would carry unique consequences for reform.

These studies were guided by a conceptual model of public response to police violence (Figure 1), which positions justification of excessive force by police and moral outrage as competing interpretive strategies with divergent consequences for support for police reform. The model specifies a theory-guided pathway in which race matters through stereotype content—particularly superhumanization—rather than as a uniform main effect. Together, these studies clarify how justification narratives align with emotional responses and shape public support for police reform.

Figure 1.

Conceptual model of public responses to police violence.



Note. Solid arrows represent directional associations with reform support; the dashed double-headed arrow indicates a negative correlation between outrage and justification as competing responses. Victim race is modeled as shaping the content of justification (e.g., racialized stereotypes such as superhumanization), rather than increasing justification overall.

Results

Study 1

The central aim of Study 1 was to examine how public justifications for excessive use of force by police related to expressions of moral outrage and calls for police accountability. We asked how justification narratives and expressions of moral outrage emerged in response to

excessive force: do they diverge from one another, and how does each connect to calls for reform? To answer this, we analyzed 257,401 YouTube comments posted across 57 widely viewed videos (mean view count ≈ 2.6 million) depicting real-world incidents of excessive force by police. These comments offered a rare window into naturalistic public discourse around police violence, including the ways in which large groups of online users justified force or demanded accountability.

Comments were classified using two complementary machine learning approaches. First, we used Google's Jigsaw Perspective API[23] to detect expressions of moral outrage (73.6% of comments; i.e., emotionally charged reactions to perceived injustice[4]). This classifier, developed by the Google and Jigsaw teams, is specifically trained to identify moral-emotional content in online discourse and has been validated for use in large-scale social media analysis. For constructs where no pre-validated classifier existed—namely, justification of force and calls for police reform—we used a GPT-3.5-based model prompted to classify each comment. Importantly, all videos analyzed in Study 1 depicted incidents of excessive force by police, as determined by independent human coders. As such, we defined justification as any comment that framed the officer's actions as appropriate or necessary despite the excessive nature of the force (19.8% of comments). In contrast, calls for reform were defined as comments that made explicit demands for police accountability or structural change (59.6% of comments). We selected GPT-3.5 for these tasks due to its strong performance on nuanced, context-dependent language classification, especially when paired with well-engineered prompting and validation protocols. Representative examples of each classification label are presented in Table 1. Multilevel logistic regression models[24]—including a random intercept for video ID—were then used to examine the relationship between justification, outrage, and accountability discourse, with each construct

binary-coded such that 0 indicated absence and 1 indicated presence. Full details on construct definitions, prompt design, classifier accuracy, and validation procedures are provided in the Methods and Supplementary Information (SI Appendix, Section 2.1).

Table 1

Representative YouTube Comments by Classifier Category

Construct	Class	Example Comment
Justification of Excessive Force	0	"Bro, why is this cop so f***ing mad??"
	1	"If you mouth off to a cop, you waive your rights. Simple as that."
Moral Outrage	0	"Wow, that spin move by the officer during the shootout was impressive."
	1	"Thugs in uniform are worse than thugs. This is absolutely disgusting."
Calls for Police Accountability	0	"So now we're charging officers just for saying slurs? That's f***ing nuts."
	1	"Draw a gun on an innocent citizen and get sent on paid vacation. Nice…"

Note. Comments are randomly drawn from a sample of 257,401 YouTube responses to videos depicting police use of force. Classifications were binary-coded, with 0 indicating the absence and 1 indicating the presence of each construct. A GPT-3.5-based model was used to classify justification of excessive force and calls for police accountability, while moral outrage was classified using the Google Jigsaw Perspective API. Comments are paraphrased and scrambled

for anonymity, and profanity was asterisked to meet publication standards. Full classifier prompts, validation procedures, and accuracy metrics are provided in the Methods and SI Appendix, Section 2.

We first tested whether the justification of excessive use of force by police was associated with expressions of moral outrage. Justification was associated with a lower likelihood of expressing outrage (OR = 0.68, 95% CI [0.67, 0.70], z = -34.42, $p < .001$), such that comments justifying police conduct were approximately 1.5 times less likely to express outrage than comments that did not. Conversely, moral outrage was strongly associated with demands for accountability: comments expressing outrage were over 7.5 times more likely to include language calling for police accountability (OR = 7.45, 95% CI [7.30, 7.61], z = 188.08, $p < .001$). Next, we examined the relationship between the justification of excessive use of force by the police and calls for police accountability. Justification was associated with a substantially lower likelihood of accountability language (OR = 0.017, 95% CI [0.016, 0.017], z = -184.47, $p <$ .001), meaning that commenters who defended police conduct were approximately 60 times less likely to call for police accountability than those who did not. These classification frequencies and effect sizes are visualized in Figure 2–which presents the overall prevalence of each comment type (Panel A) and the associated odds ratios from our regression models (Panel B).

To further examine how justification and outrage jointly shape public responses, we ran a multilevel logistic regression predicting accountability discourse with both variables entered simultaneously. Both predictors remained significant, and their magnitudes increased when modeled together: comments justifying excessive force were 91 times less likely to include calls for accountability (OR = 0.011, z = -194.37, $p < .001$), while comments expressing moral outrage were over 11.5 times more likely to include such language (OR = 11.51, z = 197.40, $p < .001$). This pattern is consistent with the idea that justification and outrage function as competing rhetorical responses in public discourse, aligning with opposing positions on accountability.

Why Reform Stalls

Finally, we explored whether the frequency of justification varied based on the race of the victim. The likelihood of justification did not differ significantly by victim race (OR = 0.84, z = −1.42, p = .155, 95% CI [0.61, 1.15]).

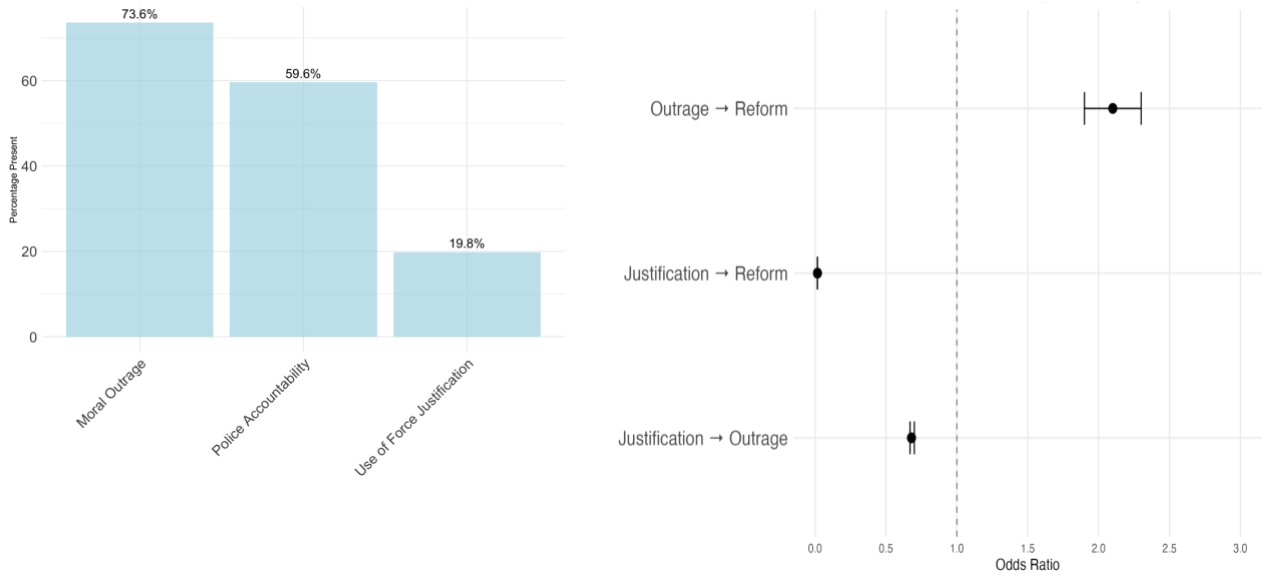
While the confidence interval included values consistent with the possibility of greater justification for Black victims, these effects were not robust, and Study 1 was not powered to detect small between-video racial differences. Moreover, our measure of justification in this study captured a broad range of narratives—blaming the victim, minimizing harm, or portraying officers as protectors—without distinguishing between different forms of justification. This makes it difficult to isolate whether specific racialized stereotypes were at play. Given the variability in how justification narratives emerge across incidents, and the likelihood that particular stereotypes map onto particular videos, broad measures may obscure race-linked patterns. One possibility is that race-linked narratives are present but not consistently expressed or detected in open comment environments, where contextual cues, video framing, and rhetorical tone vary widely. In such settings, race may not shift justification as a broad main effect, but instead shape the *form* justification takes—such as through superhumanization or dehumanization—which are context-dependent and not uniformly triggered across incidents. To build on this exploratory analysis, Study 2 tested a more specific hypothesis grounded in prior theory: whether a racialized justification portraying Black victims as superhuman—physically strong, emotionally stoic, and less vulnerable to harm—would be more likely for a Black (versus White) victim and would relate differently to outrage under controlled conditions.

Figure 2.

Justification and outrage shape public discourse on police violence.

Panel A. Proportion of comments across all videos expressing justification of excessive force, outrage, or support for reform.

Panel B. Odds ratios (ORs) linking justification and outrage to support for reform.



Note. Panel A displays the prevalence of key rhetorical responses—justification of excessive force, moral outrage, and support for police reform—across 257,401 YouTube comments. Panel B presents odds ratios from multilevel logistic regression models examining relationships between comment types. Justification was associated with significantly lower odds of outrage and reform support, while outrage predicted increased support for reform. Error bars represent 95% confidence intervals.

Study 1 provided large-scale, naturalistic evidence that justification of excessive force by police and moral outrage emerged as divergent rhetorical responses to police violence, each associated with different levels of support for police accountability. Comments that justified the officer's actions were less likely to express outrage or reference accountability, whereas comments expressing outrage were more likely to include accountability language (Fig. 2).

When both predictors were entered simultaneously, each showed a stronger association with police accountability, consistent with mutual suppression and the idea that these frames function as competing modes in public discussion. Building on this foundation, Study 2 moves from exploratory comment analysis to a controlled experimental design, directly manipulating victim race and testing whether stereotype-based justifications—particularly superhumanization—are disproportionately applied to Black victims and carry distinct implications for moral outrage and reform support.

Study 2

Building on the patterns observed in Study 1, Study 2 was designed both to replicate the emergence of justification and outrage as negatively related, competing responses with differential implications for reform, and to test a specific prediction about racialized justification. In particular, we examined whether superhumanization—a stereotype portraying Black victims as extraordinarily strong and less vulnerable—would be more likely to be endorsed for a Black (versus White) victim and whether this stereotype-based justification would carry distinct implications for reform compared to outrage. To this end, we selected two real-world incidents of police use of force—one involving a Black victim and one involving a White victim—both of which resulted in paralysis. Matching harm severity ensured that differences in responses could be attributed to victim race rather than case severity. Participants viewed each target video (order randomized), along with eight distractors, and then completed measures of justification of excessive force by police, moral outrage, and support for police reform.

A total of 159 participants (53% White, 14% Black, 51% women) were recruited via Prolific. They viewed a randomized sequence of ten short YouTube videos, including eight distractors and two target clips involving police use of force. After each target video, participants

reported their emotional reactions (including moral outrage), wrote a YouTube-style comment, and completed measures assessing their perceptions of the victim and officer (see Methods for full details). We measured participants' tendency to justify excessive force by police based on agreement with statements suggesting that the officer's actions were appropriate, that the victim was in control, or that the harm was minimal. Support for police reform was measured once at the end of the task, via agreement with seven statements reflecting willingness to engage in or endorse structural change in policing (e.g., "I am inclined to support movements calling for the defunding of the police"); internal consistency was high ($\alpha = .91$). We also examined whether participants were more likely to endorse a racialized justification—superhumanization, defined as the perception that the victim possessed extraordinary strength, control, or resilience—when the victim was Black versus White. This construct was assessed with four items adapted from the broader justification scale, designed to capture exaggerated beliefs about victim capability (see SI Appendix, section 4.1, and Methods for details).

Our analyses proceeded in three steps. First, we conducted race-agnostic regressions to test whether justification of excessive force by police and moral outrage were associated with support for police reform, providing a replication of the Study 1 pattern. Second, we examined whether these responses varied as a function of victim race, focusing on whether participants were more likely to justify force or attribute superhuman qualities when the victim was Black versus White. This step follows from our expectation that justification may be racialized through specific stereotype content, particularly superhumanization of Black victims. Third, we estimated structural equation models to test the interrelations among superhumanization, outrage, and reform. Because these variables were measured concurrently, our goal was not to adjudicate temporal order but to evaluate whether different plausible specifications captured the observed

pattern of associations. We therefore tested both a model in which superhumanization was negatively related to outrage and reform, and an alternative model in which outrage was negatively related to superhumanization and positively related to reform. Taken together, these models provide a more formal way of probing whether outrage and stereotype-based justification function as competing responses with divergent implications for reform support.

To test whether justification of excessive force by police and emotional responses were associated with support for police reform, we conducted a series of linear regression models. Greater moral outrage was significantly associated with increased support for police reform, $\beta = 0.51$, $SE = 0.09$, $t(125) = 5.58$, $p < .001$, 95% CI [0.33, 0.69]. Moral outrage was also inversely associated with justification of excessive force by police, $\beta = -0.76$, $SE = 0.06$, $t(123) = -12.69$, $p < .001$, 95% CI [−0.88, −0.64], such that participants who reported greater outrage were less likely to endorse justification rhetoric. Finally, justification of excessive force by police was significantly associated with lower support for reform, $\beta = -0.55$, $SE = 0.09$, $t(121) = -6.07$, $p < .001$, 95% CI [−0.73, −0.37]. These results were consistent with Study 1, where justification and outrage appeared to function as opposing responses, each associated with divergent levels of support for police accountability. All effects are visualized in Figure 3, which plots the standardized regression coefficients for predictors of police reform.

When both justification of excessive force by police and moral outrage were included in the same model, justification remained associated with significantly lower support for police reform, $\beta = -0.37$, $SE = 0.13$, $t(120) = -2.82$, $p = .006$, 95% CI [−0.63, −0.11]. The association between moral outrage and police reform was positive but not statistically significant, $\beta = 0.24$, $SE = 0.13$, $t(120) = 1.87$, $p = .063$, 95% CI [−0.01, 0.50]. This pattern is broadly consistent with Study 1: justification of excessive force by police showed a robust negative association with

reform support, even when emotional responses were accounted for. Although the outrage effect did not reach statistical significance in this model, its direction aligns with our theoretical expectations and suggests that shared variance with justification may have attenuated its independent association.

Before turning to the structural models, we examined whether justification and superhumanization differed by victim race. Although justification was descriptively more common for the Black victim, this difference was not statistically significant ($\beta = 0.09$, SE = 0.07, $t(126) = 1.32$, $p = .189$, 95% CI [−0.05, 0.23]). By contrast, participants were significantly more likely to attribute superhuman-like qualities to the Black victim than the White victim ($\beta = 0.19$, SE = 0.09, $t(128) = 2.14$, $p = .035$, 95% CI [0.01, 0.36]). These results suggest that while justification itself did not vary by race, racialized superhumanization may play a distinctive role in shaping how outrage translates into support for police reform, a possibility we formally test in the subsequent models.

To formally examine how superhumanization, outrage, and reform were interrelated, we estimated a set of structural equation models. The first specification showed poor fit, $\chi^2(2) = 10.06$, $p = .007$, RMSEA = .179, CFI = .951, SRMR = .079. In this specification, superhumanization was associated with lower outrage for both the Black victim ($\beta = -0.38$, SE = 0.10, $p < .001$, 95% CI [−0.58, −0.17]) and the White victim ($\beta = -0.27$, SE = 0.11, $p = .016$, 95% CI [−0.49, −0.06]). Outrage was positively associated with support for reform when the victim was Black ($\beta = 0.31$, SE = 0.14, $p = .026$, 95% CI [0.02, 0.56]) but not when the victim was White ($\beta = 0.20$, SE = 0.15, $p = .172$, 95% CI [−0.07, 0.51]). Superhumanization of the Black victim also predicted lower support for reform ($\beta = -0.27$, SE = 0.12, $p = .029$, 95% CI [−0.50, −0.03]), whereas the path for the White victim was not significant. The indirect effect of

superhumanization on reform through outrage was not statistically reliable (Black: indirect = −0.12, 95% CI [−0.26, 0.00]; White: indirect = −0.05, 95% CI [−0.17, 0.02]).

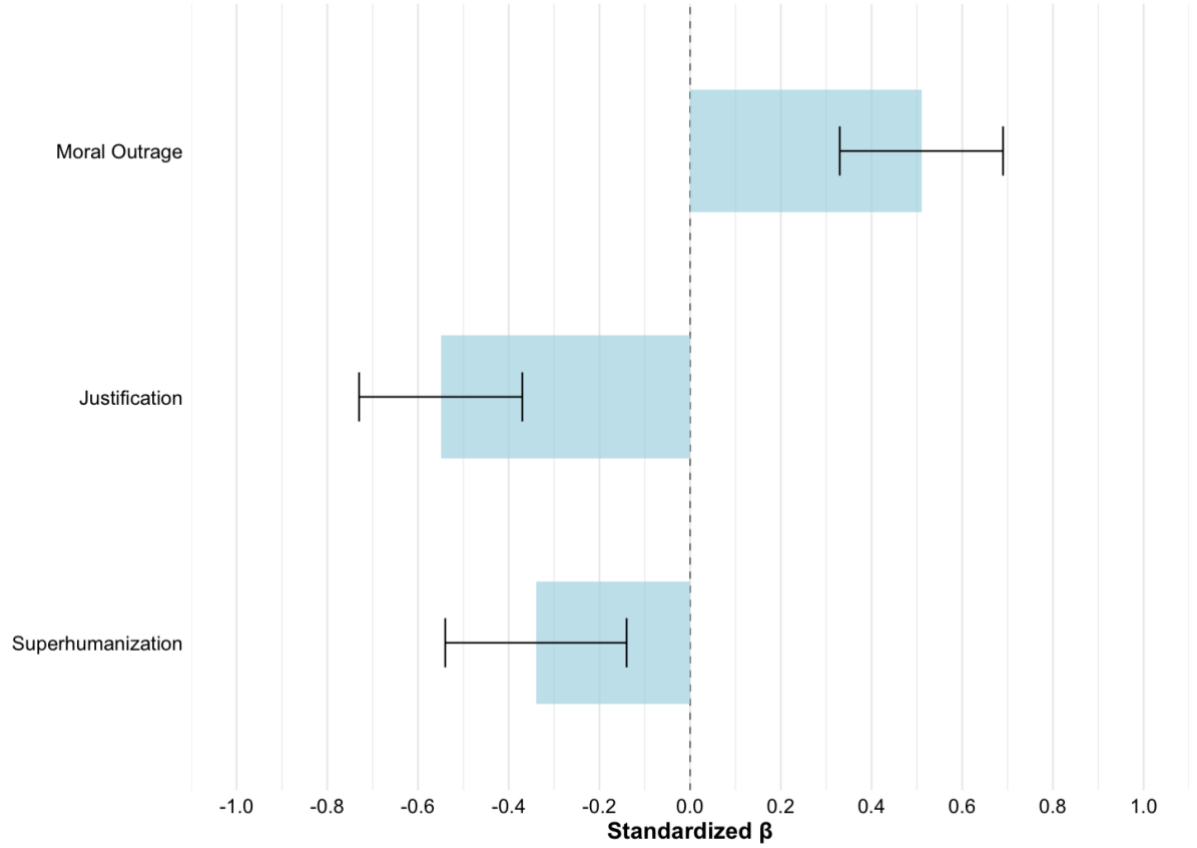
The alternative model fit the data well, $\chi^2(2) = 4.03$, $p = .133$, RMSEA = .090, CFI = .988, SRMR = .043. Outrage was associated with lower superhumanization for both the Black victim ($\beta = -0.42$, SE = 0.09, $p < .001$, 95% CI [−0.59, −0.23]) and the White victim ($\beta = -0.25$, SE = 0.09, $p = .006$, 95% CI [−0.42, −0.07]). In this model, superhumanization of the Black victim predicted lower support for reform ($\beta = -0.21$, SE = 0.12, $p = .029$, 95% CI [−0.50, −0.03]), while the corresponding path for the White victim was nonsignificant. Outrage also predicted greater reform support for the Black victim ($\beta = 0.26$, SE = 0.14, $p = .026$, 95% CI [0.02, 0.56]) but not for the White victim ($\beta = 0.16$, SE = 0.15, $p = .172$, 95% CI [−0.07, 0.51]). The indirect effect of outrage on reform via superhumanization was statistically significant for the Black victim (indirect = 0.11, 95% CI [0.01, 0.23]) but not for the White victim (indirect = −0.03, 95% CI [−0.12, 0.03]). The Black–White difference in indirect effects was also supported by the bootstrap confidence interval (difference = 0.14, 95% CI [0.01, 0.31]).

Taken together, these results replicate and extend Study 1's core finding that justification of excessive force by police and moral outrage function as negatively related, competing responses with divergent implications for reform. Justification remained a strong and consistent predictor of reduced support for reform, while moral outrage was associated with greater reform intentions. As in Study 1, the two responses also showed evidence of mutual suppression: when considered simultaneously, justification's negative link to reform sharpened, and the positive association of outrage was attenuated. Although we did not detect a significant difference in overall levels of justification by victim race, participants were more likely to attribute superhuman characteristics to the Black victim—consistent with theories of racialized stereotype

content. Structural models indicated that these racialized justifications were tied to weaker reform support and contributed to undermining the conditions under which outrage translated into reform, particularly for Black victims. This pattern highlights how stereotype-based justifications not only compete with outrage, but also shape when and how outrage mobilizes support for reform.

Figure 3.

Predictors of support for police reform in Study 2.



Note. Standardized regression coefficients (β) derived from several independent regression models predicting support for police reform. Justification of excessive force and superhumanization were associated with reduced reform support, while moral outrage was associated with increased support. Error bars represent 95% confidence intervals.

Discussion

Our findings across two studies offer novel insight into how the public interprets incidents of police violence. Across large-scale discourse and controlled experimentation, we consistently observed that justification of excessive force by police and moral outrage emerged as negatively related, competing responses with divergent consequences for reform. Justification reliably aligned with reduced support for reform, while outrage aligned with greater reform intentions. These patterns underscore that public responses to police brutality are not monolithic but bifurcate into competing rhetorical–emotional modes that either mute or amplify the call for accountability.

Study 1 revealed this divide in large-scale public discourse. Comments that justified police actions were substantially less likely to express outrage and were nearly 90 times less likely to call for police accountability, whereas comments containing outrage were over 11 times more likely to call for reform. These effects held across more than 250,000 comments from 57 widely viewed YouTube videos, and remained strong even when both predictors were modeled simultaneously. This mutual suppression pattern indicated that justification and outrage not only diverge but sharpen each other's association with reform when directly contrasted, consistent with the idea that they represent competing modes of interpretation.

Contrary to our initial expectations, the prevalence of justification did not significantly differ across videos involving Black versus White victims (OR = 0.84, z = -1.42, p = .155, 95% CI [0.61, 1.15]). While this may reflect limited statistical power to detect between-video variation—especially given our intentional focus on a smaller set of high-engagement, real-world incidents—it also raises broader theoretical questions. Specifically, it suggests that race may not

operate uniformly across public discourse, but rather through racialized narratives that are subtle, context-dependent, and may elude detection in naturalistic data without more granular or targeted measurement tools. Study 2 replicated this basic pattern under controlled conditions and extended it in three important ways. First, it confirmed that justification and outrage once again emerged as negatively related competing responses, each differentially tied to reform. Here the mutual suppression pattern was more asymmetric: justification's negative association with reform was sharpened when modeled alongside outrage, whereas outrage's positive link was attenuated. This suggests that justification may exert the more dominant pull when the two responses co-occur, underscoring its potency in neutralizing reform momentum.

Second, Study 2 allowed us to examine how race shapes justification through stereotype content. Although Study 1 did not reveal significant race differences in overall justification prevalence, participants in Study 2 were significantly more likely to attribute superhuman qualities to the Black victim than the White victim, despite both being equally harmed. This finding is consistent with prior work on superhumanization as a racialized stereotype that exaggerates strength and minimizes vulnerability. Importantly, higher levels of superhumanization were linked to both weaker reform support and reduced outrage, highlighting that stereotype-based justifications and moral outrage function as opposing responses with divergent implications for reform.

Third, we observed an unanticipated but theoretically rich asymmetry in the SEM analyses: outrage predicted reform reliably for the Black victim but not for the White victim. In race-agnostic models, outrage consistently aligned with reform, but when examined alongside justification and superhumanization, outrage carried stronger mobilizing force for the Black victim. This asymmetry may reflect the historical and systemic salience of anti-Black police

violence in the U.S., where outrage at harm to Black victims is particularly tied to collective demands for change. The implication is that racialized justifications are especially pernicious because they erode the outrage that most powerfully fuels reform.

Although our work emphasizes the corrosive role of justification, it is equally important to note that outrage was by far the more common response. In Study 1, 73% of comments expressed outrage and about 60% called for reform, whereas only about 20% offered justification. This pattern suggests a baseline public orientation toward accountability: in high-profile cases of excessive force, outrage and reform demands are the dominant responses. At the same time, our sampling strategy focused on widely viewed, highly publicized incidents where outrage may be amplified. Less visible incidents may generate more justificatory framing, underscoring the need for future research on how platform dynamics and media visibility shape response distributions.

Together, these results point to new directions for understanding how public narratives about police violence emerge and operate. One important step will be to identify the antecedents that predict whether individuals invoke justification or outrage. Just as victim race shaped the likelihood of racialized justifications, other factors such as personal stereotypes, ideology, bias awareness, or identity motives likely influence which response predominates. Conversely, certain contexts may promote outrage—for example, cues that highlight victim vulnerability, systemic injustice, or institutional betrayal. Future research should test how these antecedents and contexts channel responses toward justification versus outrage, and whether exposure to countervailing narratives can shift individuals from one response to the other. For instance, one could imagine individuals motivated toward outrage becoming less so when confronted with justificatory commentary, or individuals inclined toward justification shifting toward outrage

when presented with compelling evidence of injustice. Our data cannot adjudicate these temporal dynamics, but experimental manipulations and longitudinal designs could clarify the conditions under which these responses unfold sequentially rather than purely in competition.

Several limitations should be noted. Although Study 1 leveraged over 250,000 comments and Study 2 experimentally manipulated victim race, justification, outrage, and reform support were measured concurrently, precluding strong causal claims about temporal ordering. Our SEM analyses provide insight into the covariance structure of these responses, but model fit differences should not be interpreted as evidence of causal direction. Future work should directly manipulate justificatory narratives or emotional cues to test how they shift outrage and reform support over time. Second, although our classifiers in Study 1 performed well on validation metrics (see SI Appendix, Section 2), natural language data inevitably contain ambiguity, which likely introduced some measurement error. Such error would generally bias estimates downward rather than inflate them, suggesting that the true strength of associations between justification, outrage, and reform support may be even greater than what we observed here.

An additional limitation is that moral outrage and calls for accountability—both derived from the same set of comment classifications—may reflect similar rhetorical expressions, raising concerns about construct overlap. However, we argue that this is precisely the tension our work investigates: outrage and accountability are empirically correlated, but conceptually distinct. Outrage reflects an emotional reaction to injustice, whereas accountability represents a normative stance involving demands for punitive or structural consequences. These do not necessarily go hand in hand. In fact, one could imagine outrage serving as a performative response that fails to translate into substantive calls for change—or, conversely, individuals advocating for accountability without overt displays of outrage. Our findings help illuminate how these

rhetorical modes tend to co-occur in public discourse, particularly in relation to justification narratives.

Despite these caveats, our studies converge on a central conclusion: justification and outrage function as competing responses to police violence, each with distinct consequences for reform. Justifications—particularly those grounded in racialized stereotypes like superhumanization—dampen the outrage most strongly linked to reform support, thereby undermining momentum for accountability. Yet the dominance of outrage in our data also signals a capacity for public empathy and moral concern. Interventions to sustain reform efforts must therefore both counter justificatory narratives—such as appeals to fear, victim criminality, or racialized strength—and amplify moral outrage in ways that connect to concrete accountability demands. By clarifying how divergent responses to police violence emerge, and how racialized justifications weaken the translation of outrage into reform, our findings illuminate both the barriers and the possibilities for systemic change.

Methods

Study 1

According to the Northwestern University IRB (STU00221881), this study involving publicly available data was determined to be exempt. To examine public reactions to police violence, we collected a dataset of 57 widely viewed YouTube videos depicting incidents of police use of force. From these videos, we retrieved 257,401 unique user comments using the YouTube Data API[23] (See SI Appendix Section 1.1 for details on data collection, cleaning, and video inclusion criteria).

Why Reform Stalls

To assess how the public justifies police conduct, expresses outrage, and demands accountability, we employed two complementary machine learning approaches. First, we used OpenAI's GPT-3.5 model to classify each comment along two dimensions: (1) justification of excessive force and (2) calls for police accountability. Comments were classified as containing justification if they defended the officer's actions or blamed the victim (e.g., "he should have complied"), and as containing accountability if they criticized the police officer or institution (e.g., "these cops should be in jail"). We prompted GPT-3.5 with clear definitions and multiple examples per category to ensure reliable annotation, consistent with best practices for LLM-based classification in psychology[26]. Comments that did not engage with the incident substantively (e.g., "just watching") were coded as 0 on both dimensions. See SI Appendix Section 2 for the full classification prompt, examples, and validation metrics.

Second, we used the Google Jigsaw experimental outrage classifier to detect expressions of moral outrage in YouTube comments[23]. This classifier is not publicly available; rather, it was provided directly to the second author by the Jigsaw team for research purposes. Unlike Jigsaw's publicly available toxicity models, this experimental classifier is specifically designed to capture language associated with moral outrage—such as expressions of anger, moral condemnation, or calls for blame—rather than general toxicity or incivility. The model outputs a continuous score ranging from 0 to 1, reflecting the likelihood that a comment contains outrage. As in prior work[26], we applied a threshold of 0.5 to binarize scores into outrage present (1) or not present (0). Classifier accuracy, validation checks, and comparisons with human coding are reported in SI Appendix Section 2.

Overall, 19.8% of comments were classified as justifying police conduct, 59.6% as calling for police accountability, and 73.6% as expressing moral outrage. See Table 1 for

representative examples across categories. For model transparency and replication, we provide all prompts, thresholds, and performance metrics in the Supplementary Information. Finally, in addition to these linguistic indicators, a team of trained research assistants coded each video for victim race and additional covariates that could influence comment patterns (e.g., presence of a weapon, number of officers, outcome of the incident). Coding procedures, interrater reliability, and variable descriptions are available in SI Appendix Section 2.

Study 2

Study 2 was approved by the Northwestern University Institutional Review Board, study number STU00221383. Participants were informed during consent that they would be asked to watch YouTube videos and might experience psychological discomfort in response to some of the content. Although the true purpose of the study (assessing reactions to police violence) was withheld to preserve the authenticity of responses, participants were fully debriefed afterward and provided with detailed information about the study's aims, including access to mental health resources. To minimize distress, the selected videos depicted real incidents of police use of force but were edited to reduce identifiable features and the graphic nature of the violence where possible (see SI Appendix, section 4 for consent and debriefing materials).

The final sample consisted of 159 Prolific participants (53% White, 14% Black, 51% women; *M* age = 42.98, *SD* = 14.30), with 46% identifying as Democrat. Participants were compensated $2.50 for their time. To simulate a YouTube viewing experience, they watched a randomized sequence of ten 30-second videos. Of the ten total videos shown, eight served as filler clips designed to simulate a naturalistic YouTube browsing experience and minimize participant awareness of the study's focus. Five of these fillers featured humorous or heartwarming content (e.g., pet mishaps, surprise reunions, feel-good moments), while two

involved non-police-related instances of interpersonal conflict intended to evoke moral or emotional reactions (e.g., public road rage, a customer berating a service worker). The remaining two videos were the target stimuli—depicting real incidents of police brutality—carefully matched for severity and outcome. Both target videos involved the use of excessive force by the police that resulted in the victim being paralyzed—one involving a Black man, the other a White man. To ensure consistent titling and reduce potential source bias, the videos were re-uploaded to a controlled research account under the heading "Victim Paralyzed as a Result of Police Brutality" (see SI Appendix, Section 2.1, for selection procedures).

After each video, participants responded to a series of prompts. They first wrote a YouTube-style comment, followed by a brief reflection on their emotional reactions. They then completed a measure of current affective state. For the two police brutality videos specifically, participants rated the extent to which they perceived the victim as superhuman. After viewing all ten videos, they completed items measuring support for police reform and responded to several additional exploratory measures (see SI Appendix, Section 1).

Measures

Justification. To evaluate the extent to which participants justified the treatment of the victims, we developed a novel justification of excessive force scale composed of 11 items. This 5-point scale measured agreement with statements addressing perceptions of victim control, resilience, blame, and the appropriateness of the authorities' actions. After each police brutality video, participants rated their agreement with statements such as "The treatment of the victim was justified", "The situation was handled appropriately by the authorities involved," "The victim had control over the situation", and "The treatment of the victim by the police in the video was unfair" (reverse-coded). Items reflecting victim pain and helplessness were also reverse-coded to ensure that higher overall scores consistently reflected stronger justification of excessive force. Reliability for the full scale was high ($\alpha = .81$, $M = 2.20$, $SD = 0.60$).

Superhumanization. As an exploratory construct, we examined a four-item subset of the justification scale designed to capture attributions of extraordinary physical or emotional resilience. This subset, inspired by prior work on superhumanization[18], included statements such as "The victim possessed qualities or abilities beyond those of a typical person" and "The victim could handle more than an average person in similar circumstances". These items were averaged separately to create a superhumanization index. Reliability for this subset was satisfactory ($\alpha = .73$ for the White victim, $M = 1.73$, $SD = 0.75$; $\alpha = .75$ for the Black victim, $M = 1.86$, $SD = 0.84$).

Police Reform. To measure support for police reform, we developed a seven-item scale, each designed to assess a different facet of support for police reform initiatives, measured on a five-point agreement scale. After having observed all of the videos participants reported agreement with the following statements: "I am inclined to support movements calling for the defunding of the police", "I would actively participate in community discussions advocating for police reform", and "I would consider contributing my time or resources to support community-based policing efforts". The scale demonstrated strong internal consistency, with a reliability coefficient of $\alpha = .91$ ($M = 3.58$, $SD = 1.01$).

Outrage Responses. To measure outrage responses to each video, we used an adapted version of the *Brief Mood Introspection Scale*[28]. After watching each video, participants were asked to report the extent to which each of the 18 emotions represented their current affect (e.g., sadness, anger, shame, outrage). Additionally, for the police brutality videos, participants were asked to report the extent to which each of the 18 emotions represented the current mood of the victim. For our main analyses, we focused on the emotion of outrage experienced by the participant.

References

1. Mapping Police Violence. (2024, October 15). *Mapping Police Violence*. https://mappingpoliceviolence.org
2. Green, D. J., Duker, A., Onyeador, I. N., & Richeson, J. A. (2023). Solidarity-based collective action among third parties: The role of emotion regulation and moral outrage. *Analyses of Social Issues and Public Policy*, *23*(3), 694–723. https://doi.org/10.1111/asap.12368
3. Knab, N., & Steffens, M. C. (2021). Emotions for solidarity: The relations of moral outrage and sympathy with hierarchy-challenging and prosocial hierarchy-maintaining action intentions in support of refugees: Peace and Conflict: Journal of Peace Psychology. *Peace and Conflict: Journal of Peace Psychology*, *27*(4), 568–575. https://doi.org/10.1037/pac0000548
4. Thomas, E. F., & McGarty, C. A. (2009). The role of efficacy and moral outrage norms in creating the potential for international development activism through group-based interaction. *British Journal of Social Psychology*, *48*(1), 115–134. https://doi.org/10.1348/014466608X313774
5. Ginther, M. R., Hartsough, L. E. S., & Marois, R. (2022). Moral outrage drives the interaction of harm and culpable intent in third-party punishment decisions: Emotion. *Emotion*, *22*(4), 795–804. https://doi.org/10.1037/emo0000950
6. Chaney, C., & Robertson, R. V. (2013). Racism and Police Brutality in America. *Journal of African American Studies*, *17*(4), 480–505. https://doi.org/10.1007/s12111-013-9246-5

7. Jefferis, E., Kaminski, R., Holmes, S., & Hanley, D. (1997). The effect of a videotaped arrest on public perceptions of police use of force. *Journal of Criminal Justice*, *25*, 381–395. https://doi.org/10.1016/S0047-2352(97)00022-6
8. Hetey, R. C., & Eberhardt, J. L. (2018). The Numbers Don't Speak for Themselves: Racial Disparities and the Persistence of Inequality in the Criminal Justice System. *Current Directions in Psychological Science*, *27*(3), 183–187. https://doi.org/10.1177/0963721418763931
9. Kahn, K. B., Money, E. E. L., & Lake, J. (2024). Whose Pain Matters? Racial Differences in Perceptions of Emotional Pain After Fatal Police Shootings. *Race and Social Problems*. https://doi.org/10.1007/s12552-024-09413-1
10. Miller, S. S., Peacock, N. K., & Saucier, D. A. (2021). Propensities to make attributions to prejudice and (mis)perceptions of racism in the context of police shootings. *Personality and Individual Differences*, *182*, 111088. https://doi.org/10.1016/j.paid.2021.111088
11. Oliver, M. B., Jackson II, R. L., Moses, N. N., & Dangerfield, C. L. (2004). The Face of Crime: Viewers' Memory of Race-Related Facial Features of Individuals Pictured in the News. *Journal of Communication*, *54*(1), 88–104. https://doi.org/10.1111/j.1460-2466.2004.tb02615.x
12. Lowery, W. (2017, December 8). *Graphic video shows Daniel Shaver sobbing and begging officer for his life before 2016 shooting*. The Washington Post. https://www.washingtonpost.com/news/post-nation/wp/2017/12/08/graphic-video-shows-daniel-shaver-sobbing-and-begging-officer-for-his-life-before-2016-shooting/

13. Smith, M. (2019, April 30). *Minneapolis officer convicted of murder in fatal shooting of Justine Damond*. The New York Times. https://www.nytimes.com/2019/04/30/us/minneapolis-police-noor-verdict.html

14. Abughazaleh, K. (2023, March 14). *Fox's Martha MacCallum floats racist lie that George Floyd died of a drug overdose*. Media Matters for America. https://www.mediamatters.org/fox-news/foxs-martha-maccallum-floats-racist-lie-george-floyd-died-drug-overdose

15. Daniels, C. M. (2023, May 17). *Chauvin to ask Supreme Court to review conviction in George Floyd murder*. The Hill. https://thehill.com/regulation/court-battles/4107930-chauvin-to-ask-supreme-court-to-review-conviction-in-george-floyd-murder/

16. Bowleg, L., Boone, C. A., Holt, S. L., Del Río-González, A. M., & Mbaba, M. (2022). Beyond "heartfelt condolences": A critical take on mainstream psychology's responses to anti-Black police brutality. *American Psychologist*, *77*(3), 362–380. https://doi.org/10.1037/amp0000899

17. Lett, E., Asabor, E. N., Corbin, T., & Boatright, D. (2021). Racial inequity in fatal US police shootings, 2015–2020. *J Epidemiol Community Health*, *75*(4), 394–397. https://doi.org/10.1136/jech-2020-215097

18. Waytz, A., Hoffman, K. M., & Trawalter, S. (2014). A Superhumanization Bias in Whites' Perceptions of Blacks. *Social Psychological and Personality Science*, *6*(3), 352-359. https://doi.org/10.1177/1948550614553642 (Original work published 2015)

19. Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology*, *113*(1), 59–80. https://doi.org/10.1037/pspi0000092

20. Trawalter, S., & Hoffman, K. M. (2015). Got Pain? Racial Bias in Perceptions of Pain. *Social and Personality Psychology Compass*, *9*(3), 146–157. https://doi.org/10.1111/spc3.12161

21. Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. Journal of personality and social psychology, 94(2), 292–306. https://doi.org/10.1037/0022-3514.94.2.292

22. Glenza, J. (2014, November 25). *'I felt like a five-year-old holding on to Hulk Hogan': Darren Wilson in his own words*. The Guardian. https://www.theguardian.com/us-news/2014/nov/25/darren-wilson-testimony-ferguson-michael-brown

23. Jigsaw. (2024). Perspective API. https://www.perspectiveapi.com/

24. Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

25. Smiley, C., & Fakunle, D. (2018). From “brute” to “thug:” The demonization and criminalization of unarmed Black male victims in America. In J. J. Brunson & B. D. Miller (Eds.), Police and the unarmed Black male crisis (pp. 8–39). Routledge.

26. S. Rathje, D. Mirea, I. Sucholutsky, R. Marjieh, C.E. Robertson, & J.J. Van Bavel, GPT is an effective tool for multilingual psychological text analysis, Proc. Natl. Acad. Sci. U.S.A. 121 (34) e2308950121, https://doi.org/10.1073/pnas.2308950121 (2024).

27. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

28. Mayer, J. D., & Gaschke, Y. N. (1988). *Brief Mood Introspection Scale (BMIS)* [Database record]. APA PsycTests. https://doi.org/10.1037/t06259-000