[Prioritizing Intent in Judgments of Intergroup and Intragroup Harms]

[Daniel Yonas]

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

Abstract

[Prioritizing Intent in Judgments of Intergroup and Intragroup Harms]

[Daniel Yonas]

Researchers in moral psychology often investigate harmful behaviors at the interpersonal level but rarely consider how moral processes might play out in group contexts. Meanwhile, researchers studying intergroup processes rarely consider mechanisms rooted in moral cognition. Integrating work on moral psychology and intergroup relations, five studies examined how Black and White adults (and, in the case of Study 5, children) view intent when making moral judgments about intergroup and intragroup harms. We presented participants with a series of vignettes describing intergroup harms (e.g., a White person refusing to sit next to a Black person) or intragroup harms (e.g., a White person refusing to sit next to another White person). Participants then responded to questions about how the behaviors should be punished (Studies 1, 3, 4, and 5), the morality of the behavior (Studies 2, 4, and 5), and the degree to which the transgressor's intent mattered (all). White participants prioritized transgressors' intent more than did Black participants when evaluating intergroup harms (Studies 1, 3, and 4) and punished these harms less (Studies 2-5). These results did not emerge for intragroup harms, suggesting that White individuals may reason about harms that implicate their in-group members as perpetrators differently from harms that do not carry this implication. Taken together, these studies suggest that the difference in the emphasis placed on intent in intergroup contexts versus intragroup contexts originates in later childhood, continues into adulthood, is distinct from strict ingroup preference, and highlights the importance of considering morality and intergroup bias in concert.

Table of Contents

List of Charts, Graphs, Illustrations

Acknowledgments

"Remember no man is a failure who has friends". I do not know that I have ever resonated with these words more than now as I think back about my time in graduate school, but I can safely say that even without a degree my time here would not have been a waste because of the people that I have been fortunate enough to know. Thank you first to Larisa, who gave me the opportunity to learn more about the things I care so much about. I cannot begin to describe how much more I have grown with you as a mentor.

To each of my committee members, Tory, Nim, Melanie, and Colin. I am honored beyond words that I am able to have your feedback to make this work the absolute best that it could be.

To Jiawen, Wen, and Victoria. We started this journey together in an unprecedented times. Even if the world never really goes back to normal, I'm glad that I was able to share even a part of this journey with you. To Christian and Sophie, to me the SAMC did not start until you joined. To the rest of my Columbia family, I think I did not really understand what solidarity meant until I felt it here. To Sophie, Shelly, Hannah, Dunnigan, and all my SISPP friends, I think its these bonds that well and truly kept me going. Thank you for teaching me that researchers can and should be human first. To David, Josh, Jooney, and Janet who suffered with me and cheered my success.

To Ariel, Angel, & Jess, I could never have imagined that I would be where I am right now when I sat down in that developmental psychology class what feels like so long ago. Angel, thank you so much for pushing me to pursue this and for everything that you've done to support me as I try and make my way. Jess, you are a mentor in the fullest sense of the word. I hope that

I have been able to do you both proud. To Ariel, I've never met anyone who says "I've got you" and means it like you do.

To Avani, whose kindness, encouragement, and occasional distractions have made this journey all the more rewarding.

Dedication

To Kidan, Mailat, Hilina, and Yonas without whom I do not know who or where I would be. We made it.

Chapter 1: Introduction

In the United States, racism is often framed as a problem of individual bad actors with explicitly malicious intent. However, many racialized behaviors occur in ways that are more subtle, ambiguous, and frequently justified as unintentional. For example, in 2020, a White woman in Central Park called the police on a Black man who was birdwatching, falsely claiming that he was threatening her (Harriot, 2020). When confronted about her actions, she insisted that she "did not mean to" cause harm. This kind of reasoning—where individuals deny racist intent or use the absence of explicit racist intent to deny harm while their actions nonetheless reinforce racial inequality—reflects a broader pattern in contemporary expressions of racism (Bonilla-Silva, 2006; Bobo et al., 2012; Dovidio et al., 2000). Research on intergroup relations has increasingly focused on these covert and everyday forms of racism, such as microaggressions, racialized social avoidance, and biased institutional policies that persist despite their often-denied discriminatory impact (Freeman & Stewart, 2021; Sue & Spanierman, 2020). The current work uses the lens of moral psychology to investigate how people evaluate these behaviors and why they might downplay or excuse their harm.

Underpinnings for the importance of intent

Imagine a situation in which a Black man enters a jewelry store at his local mall near closing. The manager of the store approaches him and tells him that service workers must wait until the store is officially closed before beginning their shift. The Black man replies that he is a prospective customer. The manager, who is White, apologizes for their mistake and walks away. While the manager can leave the situation feeling a bit chagrined at their mistake, the Black man is left feeling frustrated and indignant at the assumption made by the manager that because of his race he is more likely to be a service worker than a paying customer.

This type of message is not only common but fits into a larger constellation of stereotyping and prejudice toward Black people (and other marginalized groups) more generally across a variety of contexts (Bobo et al., 2012; Bonilla-Silva, 2006; Devos & Banaji, 2005; Dovidio et al., 2002; Embrick, 2011; McCabe, 2009). Yet, if one was to interrogate the manager, they may claim that they meant no harm by their comment and that it was an honest mistake. Similarly, the Black man may believe that the manager intended no harm in their assumption about him and yet be left with his negative feelings. Alternatively, this action may have been intentionally malicious, but the man may never receive that clarity, and the nature of the intent can remain unstated even when pressed (due to a reluctance to confess to prejudicial viewpoints), leaving a lack of resolution and no accountability. Past scholarship has offered competing interpretations of situations such as this one (i.e., situations involving unintentional harm): the idea that the manager is less culpable for this negative outcome because of the lack of explicit intent (Masek, 2010; McMahan, 1994), the notion that the act is harmful regardless of the actor's intent (Kagan, 1998; Lui et al., 2020), and the related claim that the lack of explicit intent does not absolve actors of culpability for harm caused (Freeman & Stewart, 2021; Friedlaender, 2020).

Based in part on these normative claims about how much intent *should* matter, moral psychology has asked the descriptive question of how much intent *actually does* matter for laypeople's moral evaluations. Much experimental work investigating adults' moral cognition finds a sensitivity to information about intent in the process of moral judgment (Cushman, 2008; Payir & Heiphetz, 2022a; Young & Saxe, 2011). Adults reliably focus on information about intent when judging the moral wrongness or rightness of other people's behaviors (Ames & Fiske, 2013; Cushman, 2008; Heiphetz & Young, 2014). For instance, in one line of work

(Young et al., 2007), adults delivered harsher moral judgments in cases of intentional rather than accidental harm. In this study, participants evaluated scenarios in which a protagonist caused harm either intentionally (e.g., knowingly poisoning a friend) or accidentally (e.g., mistakenly believing a substance was safe when it was actually poisonous). Participants were significantly more likely to judge the protagonist as morally blameworthy when the harm was intentional, even if the outcome was the same. This work also demonstrates that people are less likely to condemn individuals who lack malicious intent, even when the consequences of their actions are equally harmful. Additionally, several studies provide evidence for a dual process model of moral judgment, whereby both harmful intent and harmful outcomes contribute to judgments of blame and punishment (Cushman, 2008; Cushman et al., 2013a; Kneer & Machery, 2019; Kneer & Skoczeń, 2023; Martin & Cushman, 2016). How individuals emphasize information about intent depends in part on what question people are answering, with people placing a higher emphasis on outcomes when assigning punishment and a higher emphasis on intent when assigning blame (Cushman et al., 2013; Kneer & Machery, 2019; Kneer & Skoczen, 2023).

With the understanding that intent plays a role in adults' judgments of harm comes the question of whether the role of intent is consistent in adults' thinking across time or may look different at different points in development. As such, developmental work in this domain often investigates how the degree to which children attune to information about intent and mental states changes across early childhood. This work finds that the early elementary school years typically mark a shift from more outcomes-focused moral judgments to more intent-focused moral judgments. There are a multitude of possible explanations for this shift including the development of theory of mind, or the influence of cultural socialization highlighting individualistic reasoning (Killen et al., 2011; Miller, 1994). In one line of work (Cushman et al.,

2013b), participants learned about one character who performed an unintentional harm (e.g., accidentally tripped and pushed someone over) and another character who failed to perform an intentional harm (e.g., tried but failed to push someone over). While 4- to 5-year-olds judged accidental harms more harshly than failed attempts at harms, older participants showed the opposite pattern. In other words, older participants prioritized intent by giving characters credit for their positive desires. In a related line of work (Killen et al., 2011), experimenters presented 3- to 8-year-olds with a story in which a well-intentioned child mistook a peer's cupcake (which was concealed in a paper bag) for trash and threw it away. Here, older children were less likely than younger children to attribute negative intent to the transgressor and to advocate for punishing a person who committed an accidental harm.

Of note, studies investigating the role of intent in children's and adults' moral judgments do not typically consider the role of group membership. For instance, prior work has asked participants to evaluate an actor who accidentally or intentionally caused someone to be sick (Young & Saxe, 2011), threw away someone else's property (Killen et al., 2011), or destroyed something that another person had built (Payir & Heiphetz, 2022a). This work focused at the interpersonal level and did not examine how the characters' or participants' group membership might shape participants' responses. However, in everyday life, harmful behaviors often occur in intergroup contexts (Sidanius & Pratto, 1999; Sue & Spanierman, 2020). Thus, the current work probed how Black and White participants evaluated instances of inter- and intra-group harm.

Considering the role of social group membership in moral judgment advances both moral psychology and the social psychology of intergroup bias. Research on morality currently suggests that neurotypical adults emphasize intent when making moral judgments, whereas autistic adults and young children place more importance on outcomes (Ames & Fiske, 2013;

Cushman et al., 2013b; Killen et al., 2011; Moran et al., 2011). However, without a consideration of group membership, this claim is incomplete. At the interpersonal level, people who have transgressed against others emphasize information about themselves (e.g., the situation they were in, their own internal processes) more than do their victims, who are more likely to emphasize the negative outcomes they experienced (Baumeister et al., 1990; Wainryb et al., 2005). Similarly, people who belong to groups that often experience group-based harm may be less likely to emphasize a perpetrator's internal processes (e.g., their innocent intent) than people who belong to groups that often inflict group-based harm. The current work tests this prediction and, in so doing, allows moral psychology to speak to a broader diversity of human experience.

In contrast to moral psychology, work on intergroup bias does investigate social processes at the group level. However, this literature does not typically draw from theorizing in moral psychology—a relevant literature in part because many morally relevant behaviors happen in group contexts (Sidanius & Pratto, 1999) and because many people perceive intergroup bias itself to be immoral (e.g., Effron et al., 2012; Gaertner & Dovidio, 2005; Siena College Research Institute, 2021). Thus, work on intergroup bias is limited in its ability to fully understand the processes shaping intergroup outcomes. The current work draws from moral psychology to test a possible mechanism—the degree to which people place importance on a perpetrator's innocent intentions when evaluating harmful behaviors—that may underlie responses to intergroup harms.

Changes in moral self-image across development

Historically, moral psychology has extensively investigated moral self-image and related concepts such as moral self-image and moral identity, recognizing their fundamental role in shaping moral judgment and behavior (Johnston & Krettenauer, 2011; Prentice et al., 2019; Sticker et al., 2021). One reason intent might matter in moral evaluations of intergroup harms in

particular is that focusing on ingroup members' innocent intentions may bolster participants' moral self-image, allowing them to maintain a positive perception of their own moral character.

Research indicates that young children as young as ages 2 or 3 years begin to develop a moral self rooted in the internalized rules and affective experiences tied to moral actions. Their moral self-image is shaped by emotions associated with compliance and transgressions, with feelings such as guilt and shame signaling moral wrongdoing (Emde et al., 1991; Kochanska, 2002). However, Kochanska's (2002) longitudinal study found that children at this age were unlikely to spontaneously describe themselves in moral terms. Instead, their conceptualization of morality was primarily shaped by compliance with parental expectations. This longitudinal study examined the early development of moral self-regulation in toddlers and preschoolers. Researchers followed children from ages 2 to 5 years, measuring their responsiveness to parental discipline, inhibitory control, and internalization of moral norms. Through observational tasks, researchers assessed how children reacted when asked to follow rules (e.g., not playing with a forbidden toy) in the absence of direct parental supervision. The study found that children who had secure attachments and experienced gentle, inductive discipline (e.g., parents explaining why certain behaviors were wrong) were more likely to exhibit early signs of moral self-regulation. These children were better able to resist temptation and demonstrated stronger internalized moral values by age 5 years, as opposed to children who experienced harsh or inconsistent discipline. The findings underscored the role of early parent-child interactions in fostering the initial stages of moral identity formation.

Despite these early moral tendencies, research has found that preschool-aged children exhibit a pronounced positivity bias—tending to overestimate their moral behavior and resist negative self-evaluations following transgressions (Harter, 1999). In experimental studies,

preschoolers and young children often failed to integrate moral failures into their self-image, demonstrating resistance to self-attributions of wrongdoing—even when they experienced negative moral emotions such as guilt and embarrassment (Boseovski, 2010; Kochanska et al., 2002; Lagattuta & Thompson, 2007). The presence of these emotions suggests that children do register their moral missteps at an affective level, but the positivity bias may function as a protective mechanism, allowing them to feel guilt or embarrassment without fully internalizing the event as reflective of their core moral identity.

As children enter middle childhood, their ability to integrate moral values into their broader self-image strengthens. Bergman (2002) conducted a longitudinal study tracking how children's moral self-image evolved across this developmental period. He found that as children aged, they coordinated their moral views and expectancies with greater consistency, showing stronger alignment between their moral reasoning and behavior. Additionally, research by Frimer and Walker (2009) suggests that children between the ages of 7 and 9 years exhibit more stable moral identities, as evidenced by increased coherence between their moral beliefs and prosocial behaviors. This developmental progression aligns with findings by Krettenauer and Johnston (2011), who demonstrated that moral identity formation becomes increasingly robust as children transition into adolescence, characterized by stronger connections between moral self-image, emotions like guilt, and the motivation to uphold moral values.

Recent research suggests that morality is not only integral to self-concept generally but may also serve as a fundamental psychological need, akin to autonomy, competence, and relatedness (Prentice et al., 2019; Söldner et al., 2024). This perspective posits that individuals are intrinsically motivated to perceive themselves as moral and to act in ways that align with moral standards, as doing so fulfills a deep-seated need for self-integrity and social belonging.

Christner et al. (2022) further argue that as children grow, moral identity is shaped by socialization processes—such as parental guidance, peer interactions, and cultural norms—that reinforce the importance of being a morally good person. These processes help children internalize moral values, integrate them into their self-concept, and develop a sense of moral responsibility that guides their behavior across contexts.

Taken together, these findings suggest that the development of moral self-image is a gradual and dynamic process, beginning in early childhood and becoming increasingly stable as children mature. This developmental trajectory has direct implications for the current research. If moral self-image plays a role in how individuals evaluate moral transgressions, then the degree to which individuals prioritize intent in intergroup harm scenarios may be linked to their moral self-image. Specifically, when observing ingroup members (individuals that share a relevant characteristic; see Tajfel et al., 1971) acting in ways that could be interpreted as prejudicial or racist, participants might perceive a threat to their moral self-image. Highlighting a potential positive aspect of these perpetrators (e.g. a possible benign intent) may enable observers to ameliorate this threat. The present work investigates this hypothesis by examining how individuals' moral self-image may shape their interpretations of perpetrator intent. In particular, we tested responses to perpetrators who committed transgressions in intergroup versus intragroup contexts to determine how group membership influences moral judgments, particularly in relation to moral self-image. By comparing these contexts, we aimed to clarify whether individuals prioritize intent differently depending on whether the harm involves their own group or another group. This distinction is crucial as moral self-image—the degree to which individuals perceive themselves as moral beings—might well influence observers' subsequent judgments, especially when their own group is implicated as perpetrators. This approach helps

bridge moral psychology with intergroup relations, offering insights into how group dynamics influence moral reasoning.

Moral self-image and intergroup harms

Though individuals engage in unethical behaviors on a daily basis (Gino et al., 2009; Hofmann et al., 2018), they also exhibit a desire to maintain a positive image of themselves for their own benefit as well as that of others (Adler, 2006; Aquino & Reed, 2002). This desire extends to how others morally perceive them (i.e., people want others to perceive them as morally good) as well as how they morally perceive themselves (i.e., people want to perceive themselves as morally good; Dunning, 2007; Monin & Jordan, 2009). Further, this desire motivates individuals to act in ways to maintain positive moral self-image when they behave in ways that are contrary to positive perception (Barkan et al., 2012; Monin & Jordan, 2009). After engaging in immoral behaviors, individuals perceive a potential threat to their moral self-image and seek to buffer it by engaging in moral actions (Jordan et al., 2011; Sachdeva et al., 2009). One specific study (Mulder & Aquino, 2013) found that individuals who are high in moral self-image, meaning that they believed themselves to be particularly good and virtuous, were more likely than individuals with low moral self-image to donate more to charity after lying or engaging in other deceitful behavior. These findings suggest that the potential difference between an individual's moral self-image and the reality of that individual's morally relevant behavior presents a psychological pressure that leads individuals to engage in behaviors that aligns with their internal beliefs about their moral character or rationalize the disparity between perception and reality (Jordan et al., 2015).

Given these findings, and given that morally relevant behaviors often occur in intergroup contexts (Gaertner & Dovidio, 2005; Sidanius & Pratto, 1999; Wohl et al., 2006), it is

worthwhile to examine how individuals might react when perceiving a threat to their moral self-image due to the actions of their ingroup members. It is possible individuals who observe (or read about) ingroup members that harm others will experience a threat to their own moral self-image (Wohl et al., 2006). This threat may engender a similar psychological pressure to maintain a positive moral image. Past research corroborates the degree to which individuals experience this threat to moral self-image and the tendency to engage in compensatory behaviors to maintain a positive image of themselves and their group (Steele et al., 2002; Tajfel et al., 1971). Compensatory behavior here refers to behaviors that might help to repair damage to self-image by acting in such a way as to highlight or display the depth of the quality under threat. One study by Leach, Ellemers, and Barreto (2007) explicitly examined the centrality of morality to group identity. In this study, participants evaluated their ingroups and outgroups on dimensions of morality, competence, and sociability. The researchers found that morality was the most critical factor in the positive evaluation of ingroups, surpassing competence and sociability. When participants perceived their ingroup as lacking in morality, they experienced a threat to their group identity and were motivated to restore a positive moral image, often by emphasizing the moral virtues of their ingroup or engaging in behaviors that reaffirmed their group's moral standing. Thus a threat to group morality can directly influence individuals' efforts to restore a positive moral self-image, both for themselves and their group, in a manner similar to a threat to individual moral self-identity.

Another line of work more directly investigated the potential influence of group threat in the context of historical atrocities committed against indigenous people in Sweden (Čehajić-Clancy & Olsson, 2024). Before completing the main body of the study, participants read a one-page-long text documenting historic discrimination against Sami people (a group indigenous to

Sweden) representing a threat to their group identity. Participants then read stories about moral exemplars and completed measures on moral elevation (the degree to which they were inspired to behave in moral ways), both individual and group moral identity (the degree to which they perceived themselves and Swedish people generally to be moral), their moral emotions (e.g., guilt), and endorsement of reparations. Across several studies, Čehajić-Clancy and colleagues found a strong relationship between moral identity and support for reparations benefiting former victim group members. This relationship remained significant even after accounting for group-based guilt. Additionally, the studies revealed that individual moral identity—and the perceived threat to it—was a stronger predictor of compensatory behavior than group-based identity (Čehajić-Clancy & Olsson, 2024).

To return to the example described in the introduction, one potential explanation for the dynamic at play is that White individuals may experience and react to threats to their moral self-image related to societal inequalities that privilege their racial group. Knowles et al. (2014) theorize that White participants engage in three types of actions to buffer their moral self-image against threats imposed by belonging to a group that perpetuates racial harms: denying the existence of racial privilege (Lowery et al., 2007), distancing themselves from the identity that connects them to the threat after experiencing a challenge (Chow et al., 2008), and dismantling the systems that generate these inequalities by accepting the existence of privilege and endorsing policies aimed at reducing inequality (Knowles et al., 2014). The current work extends this framework by proposing a fourth strategy: *diffusing*. Rather than denying, distancing, or dismantling, *diffusing* involves emphasizing morally positive characteristics of the ingroup—such as benevolent intentions—in order to minimize the moral weight of transgressive behavior. This strategy allows individuals to acknowledge that harm occurred without fully internalizing

its moral implications, thereby preserving both a positive moral self-image and ingroup affiliation. This dissertation might find that intent-focused moral reasoning may function as a form of diffusing in response to perceived threats to moral identity, particularly when evaluating ingroup members who commit racially harmful acts.

To apply this model to the context of the present work, if participants observe ingroup members transgressing, participants may be motivated to highlight something good about those transgressors (e.g., their positive intent) as a way to indirectly maintain their own moral self-image. This process is distinct from self-affirmation as it occurs in the exact same domain the threat originated from. However, in cases where participants belong to a group that experiences harm, no effort to maintain positive moral self-image is necessary because their group is not the transgressor; in fact, people often judge victims as morally good just by virtue of being victims (Gray & Wegner, 2011; Jordan & Kouchaki, 2021). Instead, if participants belong to a group that is on the receiving end of harm, they may emphasize outcomes more because that is what they experience. To return to the previous example, the Black man might focus more on the harm originating from the store manager's assumption because he experiences these types of harms more often over the course of his life and cannot investigate the intent underlying each instance of harm. On the other side, White individuals who do not commonly experience these harms may then prioritize intent because a focus on intent does not represent the same load. Taken together, these points suggest that White and Black participants may provide different moral judgments of unintentional intergroup harms, a hypothesis tested in the current work.

The present research

The present research leverages insights from moral psychology and the social science of intergroup bias to better understand how children and adults think about intent in the context of

intergroup harms. Addressing this question provides a more complete picture of how moral judgment varies across social groups and of the mechanisms underlying potential group-based differences in moral evaluations. Critically, work in social psychology interested in how individuals form a common group identity finds that individuals identify more with others that are similar to themselves but are also capable of forming these group associations on arbitrary characteristics (e.g., T-shirt color). Both children and adults form arbitrary groups with others based on meaningless characteristics (e.g., art preferences; Tajfel et al., 1971) even when the only knowledge of these others is digital, imagined, or second-hand (through videos or vignettes; Isbell & Tyler, 2003; Janneck et al., 2013). As such, vignettes may be viable vehicles for adults to form common group identities, especially when the characteristics held in common are meaningful like race and gender (Dunham et al., 2011; Mummendey et al., 2000; Schug et al., 2013; Valdesolo & DeSteno, 2007). Thus, the following studies leveraged vignettes to investigate how Black and White individuals make moral judgments of intergroup and intragroup harms.

Study 1 probed the question of how Black and White adults think about both of these types of harms and tested the extent to which adults' moral judgments of intergroup and intragroup harms depended on the importance participants placed on perpetrator intent. Study 2 investigated Black and White adults' punishment judgments in the intergroup and intragroup contexts highlighted in Study 1 and investigated the possible connections between punishment judgments and intent. Study 3 examined how Black and White individuals' punishment judgments may differ in intergroup contexts where malicious intent was explicit. Study 4 tested whether a desire to protect moral self-image underpinned the differences in Black and White individuals' prioritization of intent when responding to intergroup and intragroup harms. Finally,

Study 5 examined how potential differences in individuals' judgments of intergroup and intragroup harms vary across age, as the way people conceptualize intent and consider their morality to be central to their overall identity (i.e. moral identity) changes across development. Together these studies provide insight into how the process of moral cognition shifts when considering the intergroup nature of the harm and group membership of the participants.

Chapter 2: Study 1

Study 1 investigated differences between Black and White adults with regards to how they make moral judgments about intergroup and intragroup harms. Additionally, this study investigated the degree to which Black and White adults emphasized intent when making these judgments. In order to gain deeper insight into participants' judgments, we collected qualitative data about the reasons behind their responses in addition to quantitative judgments probing the perceived morality of perpetrators' behaviors and how important the perpetrator's intent was to participants' evaluations. Much of the work in moral psychology is quantitative and leveraging the strengths of qualitative data (e.g. capturing factors other than intent that participants may attune to when making moral judgments) allowed us to form a more complete picture of the way participants took intent into consideration when making their punishment judgments. We pre-registered this study at https://aspredicted.org/JRP_N7P.

2.1 Method

2.1.1. Participants. Our final sample included 81 White adults (M_{age}=37.60 years, SD_{age}=14.77 years; 58 female, 20 male, 3 other) and 85 Black adults (M_{age}=38.95 years, SD_{age}=15.39 years; 55 female, 27 male, 3 other). We determined sample sizes prior to collecting data using G*Power, indicating 80% power for a two-tailed *t*-test and estimating an effect of Cohen's *d*=.50. In addition to the adults described above, we tested and excluded 24 other adults from these analyses for failing to answer attention check questions (that asked participants to briefly summarize 2 of the vignettes that they viewed) correctly.[1] We recruited participants via Prolific, collected their data through Qualtrics, and compensated them at a rate of $10/hr.

[1]Including all respondents in analyses yielded similar patterns of results as those reported in the main text, with the following exception: the significant effect of group on intent emphasis became non-significant (*p*=.439) due to the loss of power.

2.1.2. Procedure. We presented participants with a series of vignettes depicting two types of situations: cases where White individuals interacted with Black individuals in ways that could be read as intergroup harms (e.g., a White person moved to a different place on the bus when a Black couple sat next to them), and intragroup cases where White individuals performed similar behaviors toward other White individuals (e.g., a White person moved to a different place on the bus when a White couple sat next to them). After reading each vignette, participants first answered a series of counterbalanced questions that investigated their moral judgments (e.g., “In your mind, how wrong or right is what [agent] did?”, “In your mind, how good or bad is what [agent] did?”, “In your mind, how mean or nice is what [agent] did?”, and after the final moral question, “In a few words, please describe why you answered the way you did? For instance, why did you think the [agent’s] behavior was wrong or right?”). These questions were coded on a scale where 1 represented the most negative judgment (e.g., very morally wrong) and 7 represented the most positive judgment (e.g., very morally right). In the second block, participants read each vignette again. This time, they indicated the importance of intent to their moral judgment on a scale from 1 being not at all important to 4 being neither unimportant nor important to 7 being very important. These questions (e.g. "How important was what [agent] meant to do / what [agent] was trying to do / what [agent's] intentions were?") investigated the degree to which different types of harms prompted participants to emphasize intent differently.

We also coded open-ended responses for references to intent as well as other themes that emerged during coding. First, we explored participants' responses for prospective categories relevant to moral judgment. Then, we created a codebook detailing criteria that delineated one category from another (e.g. responses that included explicit references to intent and examples of what those responses would look like). Next, two different researchers coded a subset of the data

for the presence of all possible themes, entering a score of 0 if a given theme was not present in a specific participant's response and a score of 1 if that theme was present. Researchers discussed differences in their coding to reach agreement on any items where they disagreed. After aligning their coding in this way, the two researchers coded the rest of the dataset separately and tested for reliability (Cohen's κ =0.75).

2.2 Results

To determine how moral evaluations of intergroup and intragroup harms might vary across participants from different social groups, we averaged participants' responses to the quantitative moral judgment questions described above and entered this score into a 2 (Participant Race: Black vs White) x 2 (Type of Harm: Intergroup vs Intragroup) mixed ANOVA with repeated measures on the second factor. This analysis revealed a main effect of Type of Harm (F(1,164)=441.11, p=.001, partial η_p^2=.73): overall, participants reported that intergroup harms (m=3.29, sd=.62) were less morally right than intragroup harms (m=4.57, sd=.60). The effect of Race (F(1,164)=1.074, p=.302, *partial* η^2=.02) and the Race x Type of Harm interaction (F(1,164)=2.974, p=.086, *partial* η^2=.02) did not reach significance.

To determine whether the emphasis that participants placed on intent varied depending on the participant's social group membership, we averaged participants' responses to the quantitative intent questions described above and entered this score into a 2 (Participant Race: Black vs White) x 2 (Type of Harm: Intergroup vs Intragroup) mixed ANOVA with repeated measures on the second factor. This analysis revealed a main effect of Type of Harm (F(1,164)=40.24, p=.001, *partial* η^2=.20), a main effect of Race (F(1,164)=4.64, p=.033, *partial* η^2=.03), and a Participant Race x Harm Type interaction (F(1,164)=11.60, p=.001, *partial* η^2=.07; Fig. 1). To further investigate this interaction, we conducted two sets of tests. First, we compared the extent

to which Black versus White participants prioritized intent. White participants (m=5.59, sd=1.03) prioritized intent for intergroup harms more than did Black participants (m=5.06, sd=.69, F(1,84)=9.28, p=.003, CI[.184, .863]), but this effect was not present for intragroup harms (F(1,164)=0.00, p=.993, CI[-.216, 2.15)]. Second, we compared the extent to which Black participants and, separately, White participants prioritized intent for intergroup versus intragroup harms. Members of both racial groups prioritized intent for intergroup harms more than for intragroup harms (Black participants: $m_{intergroup}$=5.06, $sd_{intergroup}$=1.18, $m_{intragroup}$=4.84, $sd_{intragroup}$=.71, t(1,84)=39.65, p<.001, CI[4.808,5.315]; White participants: $m_{intergroup}$=5.59, $sd_{intergroup}$=1.03, $m_{intragroup}$=4.84, $sd_{intragroup}$=.69, t(1,80)=48.89, p<.001, CI[5.358,5.812]). However, the effect size was larger for White participants (Cohen's d=.86) than for Black participants (Cohen's d=.23).

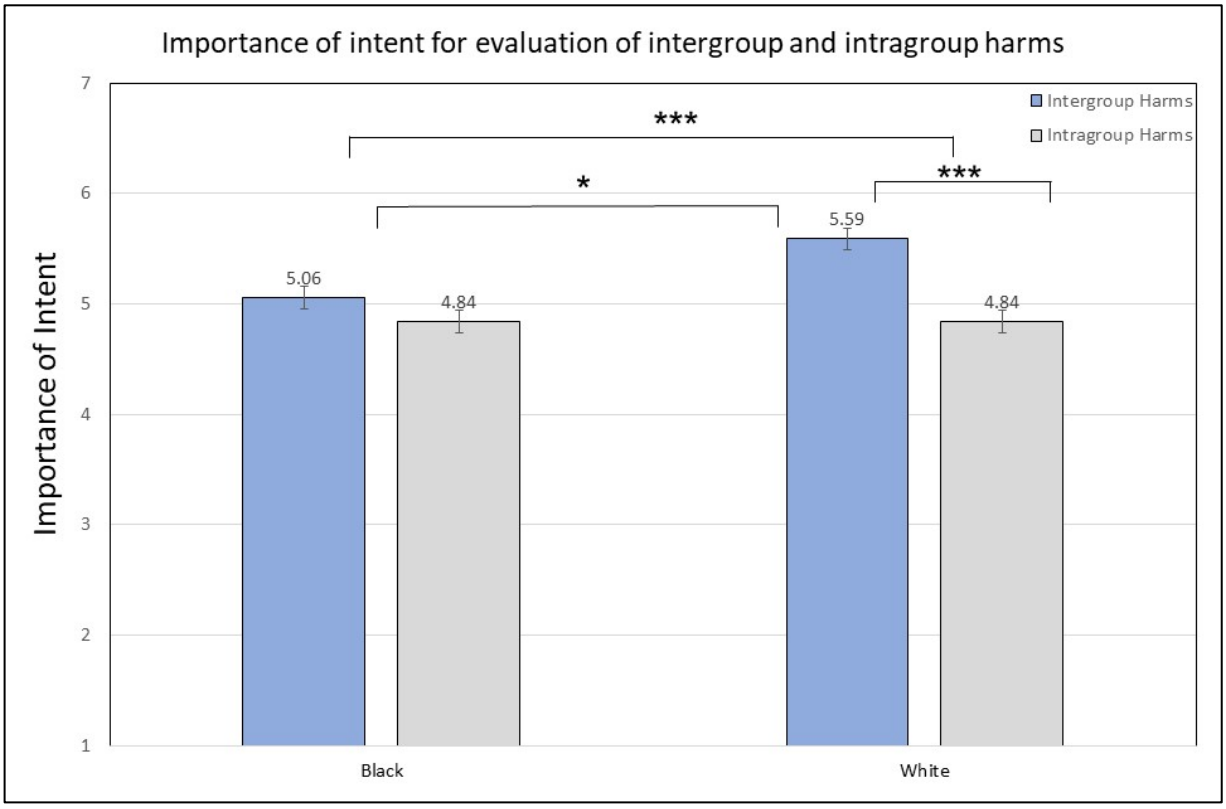

Fig 1. Black and White adults' responses about the importance of intent in the process of making moral judgments about Intergroup and Intragroup harms, Study 1. Error bars represent 95% confidence intervals.

To gain insight into the relationship between participants' moral judgments and the degree to which they emphasized perpetrator intent, we first correlated these two variables. The more participants prioritized intent, the more positive were their moral judgments of intragroup harms ($r(164)=.66$, $p=.001$), but this effect was not present for intergroup harms ($r(164)=.06$, $p=.47$). We next tested the extent to which the emphasis participants placed on intent mediated the relationship between the type of harm about which participants responded (intergroup versus intragroup) and participants' moral judgments of that harm.[2] To do so, we conducted a mediation analysis with 5,000 bootstrapped samples (Hayes, 2017). This mediation included Type of Harm (0=Intragroup, 1=Intergroup) as the predictor variable, the degree to which participants responded that intent was important in their moral judgments (intent emphasis) as a mediator, and participants' moral judgments as the outcome variable. This mediation model revealed a significant indirect effect for both Black participants (B =-.23, p=.001, 95% CI[-.176, -.281]) and White participants (B=.06, p=.020, 95% CI[.009, 123]; Figs. 2 and 3). Considering intent in the model reduced the negativity of the direct effect. That is, participants perceived intergroup harms as morally worse than intragroup harms, but considering intent suppressed a portion of this

[2] As described in the pre-registration, we also examined participants' emphasis on intent when making moral judgments as a potential mediator between participant race and moral judgments of intergroup harms. This mediation model revealed no significant indirect effect (B=.02, 95% CI[-.080,.022]). However, the mediation presented in the main text more directly addresses Study 1's substantive research question regarding the extent to which intent-based reasoning underlies the difference in moral judgment regarding intergroup versus intragroup harms.

negativity. Considering the intent of a person who engaged in intergroup harm may therefore reduce the harshness with which participants evaluate that harm.

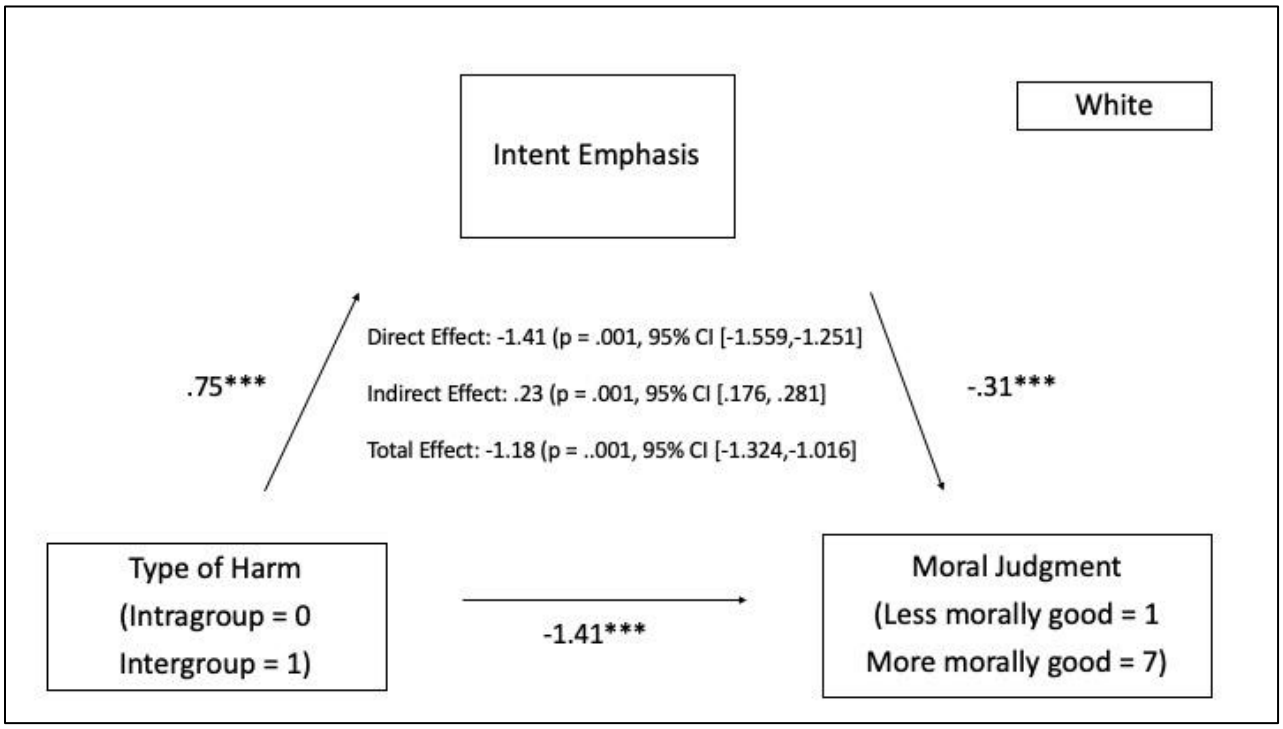

Fig 2. Mediation analysis for the relationship between type of harm, intent emphasis, and moral judgments for White participants, Study 1.

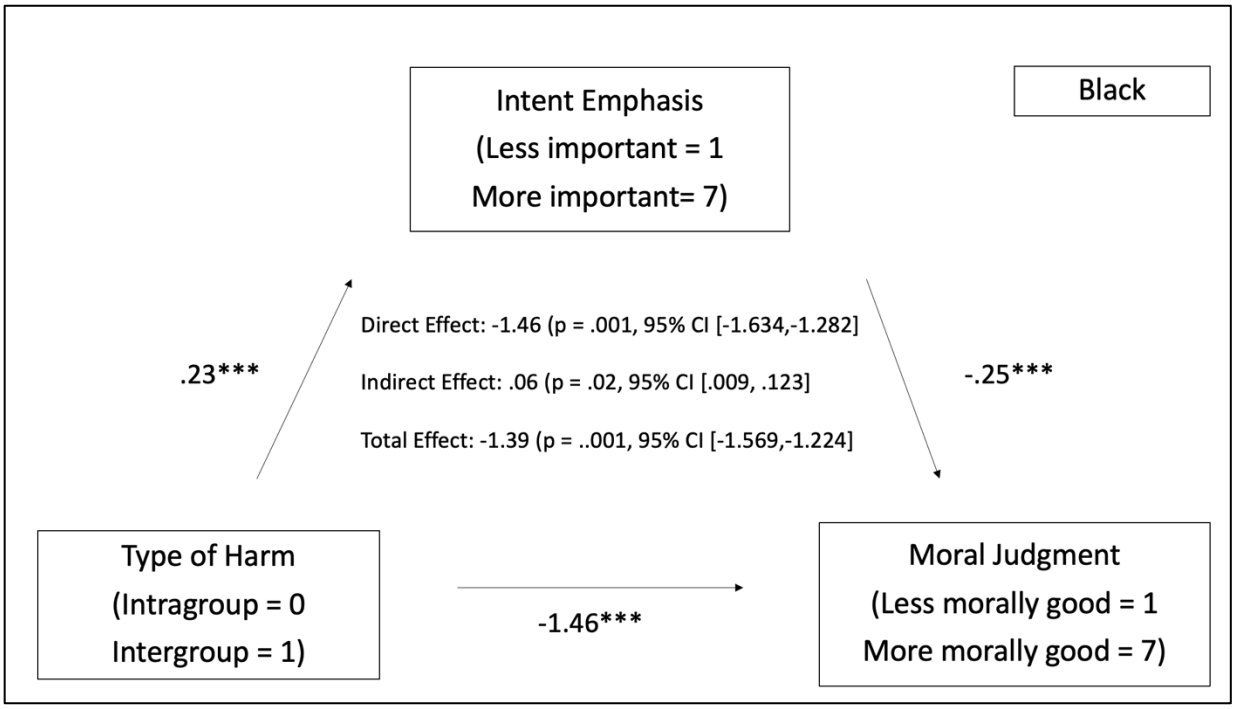

Fig 3. Mediation analysis for the relationship between type of harm, intent emphasis, and moral judgments for Black participants, Study 1.

We also coded open-ended responses for references to intent as well as other themes that emerged during coding (for example, *environmental concerns* marked instances where participants' explanations referenced the outcome of specific behaviors for individuals in the vicinity of the harm; as another example, *lack of perceived harm* marked instances where participants did not believe a harm deserving of consequence had occurred). We assigned a score of 1 to responses that referenced intent or the desire to engage in the specific behavior or cause the harm as justification for the moral judgment, whereas a lack of reference to intent or the internal mental state of the perpetrator received a 0. Below are examples of responses that would receive a code of 0 versus 1.

0 (no reference to intent): "*He treated the man like a criminal. That's wrong!"*

1 (references intent): *"I am not very clear on what the guard's intentions were in this instance. He could have been trying to see what was inside of the vehicle but there was no reason to do that in addition to questioning them"*

We entered these codes as a mediator in a model where type of harm served as the predictor variable and moral judgments, averaged across the three items above, served as the outcome variable. Here we found no significant indirect effect (*B* =-.001, 95% CI[-.048, -.043]) for Black or White participants (*B* =-.001, 95% CI[-.060, .055]).

2.3 Discussion

Study 1 revealed important differences in how Black and White adults make moral judgments about intergroup and intragroup harms, particularly in how much they prioritize intent in their evaluations. While both groups considered intergroup harms more morally wrong than intragroup harms, the emphasis on intent on was greater for White participants as compared to Black participants, and this difference was more pronounced for intergroup harms. This finding highlights that the process of moral judgment can look different based on the context of the harm being evaluated. The qualitative data provided deeper insight into these judgments, showing that individuals do make specific reference to intent in their moral evaluations, but this did not necessarily predict individuals' moral evaluations. However, moral judgments represent only one part of the wide spectrum of social judgments that adults make when faced with harms in group contexts. Adults must also make decisions about if and how these types of harms should be punished, even in contexts where intent is less clear. Thus, Study 2 asked how judgments about intent may shape punishment decisions.

Chapter 3: Study 2

Study 1 found a significant difference such that White participants emphasized intent more when evaluating ambiguous intergroup harms than did Black participants but did not find a significant difference with regards to White versus Black participants' overall moral judgments. Study 2 probed how punishment judgments (unlike the moral judgments tested in Study 1) may change in contexts where explicit intent is not present, as punishment represents a relevant consequence for differences in intent emphasis. Information about intent is especially valuable when making judgments about punishment and culpability (Cushman, 2013). This research argues for a two-process model that is at play in adults' evaluations of accidental harms. This model suggests that adults prioritize intent differently depending on whether they are making punishment judgments versus moral judgments such that mental states like intent are more important in judgments of punishment and blameworthiness . Therefore, Study 2 built on Study 1 by investigating punishment decisions rather than moral evaluations. We preregistered this study at https://aspredicted.org/JRP_N7P.

3.1 Method

3.1.1. Participants. Our final sample included 130 White adults (M_{age}=40.13 years, SD_{age}=14.20 years; 80 female, 42 male, 8 other) and 120 Black adults (M_{age}=37.00 years, SD_{age}=13.31 years; 58 female, 60 male, 2 other). We determined sample sizes prior to collecting data using G*Power, indicating 80% power for a two-tailed *t*-test and estimating an effect of Cohen's *d*=.50. We recruited participants via Prolific and compensated them at a rate of $10 per hour for their participation in a 10-minute study. In this study we did not recruit participants who had participated in Study 1. In addition to the participants described above, we tested an

additional 40 adults that we excluded from the analyses for failing to answer attention check questions.[3]

3.1.2. Procedure. We asked participants to read vignettes identical to Study 1 and then respond to multiple punishment questions and then answered multiple punishment questions probing the degree to which the instigating agents in the vignettes should face repercussions for their actions on a scale where 1 equaled very little potential punishment and 7 indicated a great deal (e.g., "In your mind, how much should the [agent] be punished for what they did?", "In your mind, how severe should the consequences be for what the [agent] did?", "In your mind, how much trouble should the [agent] be in?", and after the final punishment question, "In a few words, please describe why you answered the way you did?"). Like in Study 1, participants also justified their responses after each individual vignette. After this first block, participants answered the same intent measures described in Study 1.

3.2 Results

Similar to Study 1, we averaged participants' responses to the punishment and intent emphasis questions described above and entered these averaged responses into a 2 (Participant Race: Black vs White) x 2 (Type of Harm: Intergroup vs Intragroup) mixed ANOVA with repeated measures on the second factor.

Unlike Study 1, intent analyses revealed no main effect of Participant Race ($F(1,248)=.044$, $p=.833$, *partial* $\eta^{2=}.00$) or Type of Harm ($F(1,248)=2.07$, p=.152, *partial* $\eta^{2=}.008$). However, we did observe a significant interaction between Participant Race and Type of Harm ($F(1,248)=8.201$, p=.005, *partial* $\eta^{2=}.032$). To break down this interaction, we conducted two sets of tests. First, we investigated how White and, separately, Black participants

[3] Including all respondents in analyses yielded similar patterns as those reported in the main text.

reasoned about intergroup versus intragroup harms. This analysis revealed that White individuals prioritized intent more for intergroup harms (m=4.69, sd=1.24) than intragroup harms (m=4.76, sd=1.46; t(247.74)=-2.095, p=.037). However, Black individuals did not differ (t(237.95)=.602, p=.548) with respect to how they prioritized intent for intergroup (m=4.84, sd=1.48) or intragroup (m=4.96, sd=1.50) harms. Second, we compared how White versus Black participants evaluated intergroup harm and, separately, intragroup harm. Black and White individuals did not differ with regards to how they prioritized intent for intragroup harms (t(243.89)=-1.037, p=.301) or intergroup harms (t(233.03)=1.584, p=.115).

The punishment analyses revealed a significant main effect of Participant Race (F(1,248)=11.50, p=.001, *partial* $\eta^{2=}$.044, Fig. 4): overall, White individuals (m=1.82, sd=.71) punished all harms less severely than did Black individuals (m=2.15, sd=.78). We also observed a significant effect of Type of Harm (F(1,248)=8.20, p=.005, *partial* $\eta^{2=}$.032): participants punished intergroup harms (m=2.42, sd=1.02) less severely than intragroup harms (m=2.80, sd=.82). However, the Type of Harm x Participant Race interaction effect did not reach significance (F(1,248)=2.747, p=.099, *partial* $\eta^{2=}$.011).

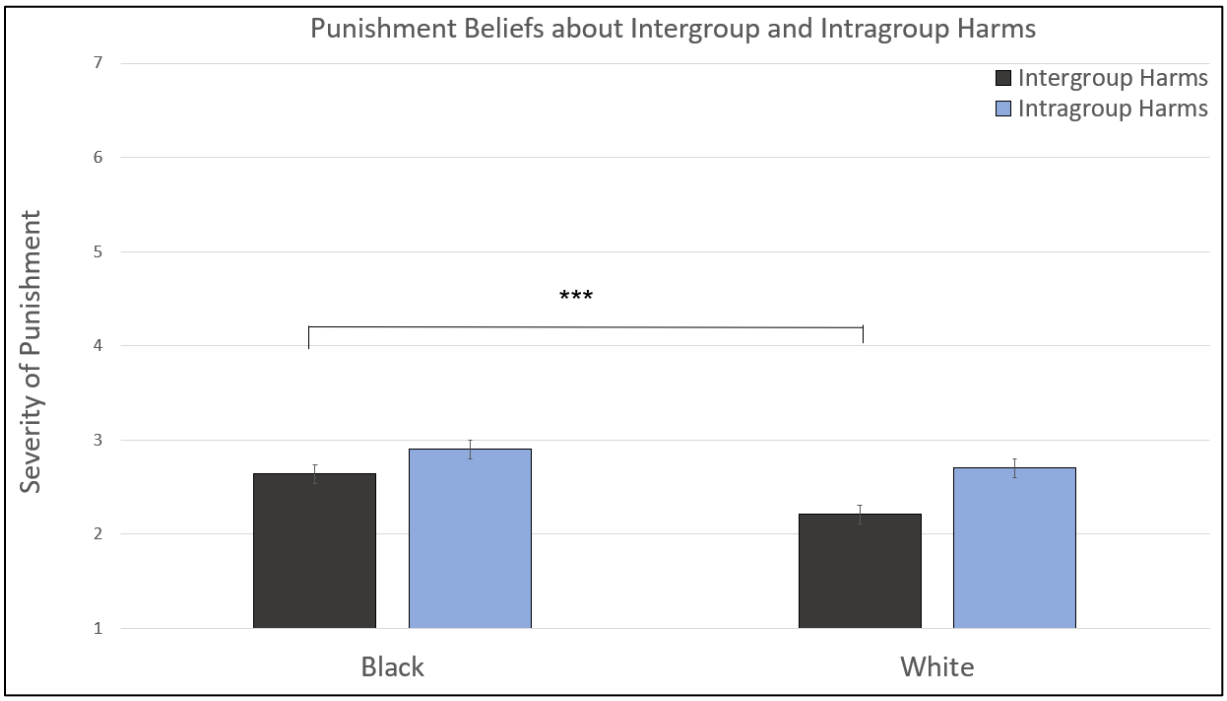

Fig 4. Black and White adults' judgments about the severity of punishment that intergroup and intragroup harms should receive, Study 2. Error bars represent 95% confidence intervals.

Similarly to Study 1, we conducted a mediation analysis with 5,000 bootstrapped samples.[4] This mediation included Type of Harm (0=Intragroup, 1=Intergroup) as the predictor variable, emphasis on intent as the mediator, and punishment decisions as the outcome variable. This mediation model did not reveal significant indirect effects for either Black participants (*B*=-.001, *p*=.800, 95% CI[-.001,.001]) or White participants (*B*=.001, *p*=.920, 95% CI[-.003,.001]).

[4] As in Study 1, the pre-registered mediation included Participant Race (0=White, 1=Black) as the predictor variable, emphasis on intent for intergroup harms as the mediator, and punishment beliefs about intergroup harms as the outcome variable. This mediation model revealed no significant indirect effect (*B*=.02, 95% CI[-.080,.022]).

Finally, we coded open-ended responses for references to intent as well as other themes that emerged during coding. We entered these codes as a mediator in a model where Participant Race served as the predictor variable and punishment judgments of intergroup harms, averaged across the three items above, served as the outcome variable. Again, we found no significant indirect effect (*B* =-.042, *p*=.673, 95% CI[-.128, .039]).

3.3 Discussion

Study 2 built on Study 1 by investigating Black and White participants' punishment decisions in intergroup and intragroup contexts. Unlike Study 1, Study 2's intent analyses revealed no overall differences based on participant race or type of harm. However, there was a significant interaction between participant race and type of harm. Breaking down this interaction, we found that White participants prioritized intent more for intergroup harms than for intragroup harms, whereas Black participants did not differ in how they prioritized intent across contexts. These findings suggest that while White participants may adjust their emphasis on intent depending on the context of harm, Black participants' prioritization of intent remains consistent across contexts. Further, this study showed that, collapsing across contexts, Black participants expressed more punitiveness than did White participants. Nevertheless, both Black and White participants seemed to align in their belief that intragroup harms require more severe punishment than their intergroup counterparts. These results present an interesting pattern in conjunction with Study 1, in which both Black and White participants reported that intergroup harms were more morally wrong than intragroup harms. Why would individuals perceive intergroup harms as more morally wrong than intragroup harms, yet punish these harms less? One possible explanation is that all participants are prioritizing the perspective of the agent when imagining a situation where they are accidentally racist toward another person. While such behavior is clearly

wrong, the perceived commonality of this type of harm may lead participants to view it as less deserving of severe punishment and or feel less comfortable assigning severe punishment on behalf of another group. Further, while participants can use punishment in order to correct the behavior of perpetrators, it is possible that participants perceive that these types of intergroup harms might step from societal level issues, making individual punishments for ambiguously intentional harms less meaningful. Another possibility is that although participants may have learned that it is wrong for White individuals to harm Black individuals, they may nevertheless place greater value on harms experienced by White rather than Black individuals (Dovidio et al., 2002; Eberhardt et al., 2006; Norton & Sommers, 2011). Because intragroup harms in Studies 1 and 2 always involved White victims, participants may have reported that more severe punishment was appropriate than in intergroup cases, which always involved Black victims (see also Eberhardt et al., 2006; Medwed, 2020). Study 4 further took up this question by investigating participants' judgments of multiple types of intergroup harms, including situations where White people harmed Black people and situations in which Black people harmed White people.

Although Studies 1 and 2 used different dependent measures, they both measured the extent to which participants emphasized intent when making morally relevant decisions. One benefit of this approach is that it revealed information about naturally occurring variation in Black and White participants' judgments regarding intent. However, one drawback of this correlational approach is that it does not allow for conclusions about any causal role that a transgressor's intent may play in participants' moral decision-making. Thus, Study 3 extended Studies 1-2 by manipulating information about intent.

Chapter 4: Study 3

Study 1 found that both White and Black participants evaluated intergroup harms as more wrong than intragroup harms. Further, when reasoning about intergroup harms, White participants placed more importance on the perpetrator's intent than did Black participants. Study 2 revealed that White participants prioritized intent more for intergroup harms than for intragroup harms, whereas Black participants did not differ in how they prioritized intent across contexts. Additionally, there were no differences between Black and White participants in how they prioritized intent for either intragroup or intergroup harms when compared directly. Further, Study 2 found that Black and White individuals punished intergroup harms less severely than intragroup harms. While these two studies adopted a correlational approach to measure the importance participants placed on intent, Study 3 investigated the causal role of intent in moral judgment by manipulating information about this factor. That is, Study 3's vignettes explicitly stated whether or not the perpetrator meant to perform a harmful action. Because Studies 1 and 2 revealed significantly different judgments across race in the intergroup condition, this was the only condition tested in Study 3 (we removed the intragroup condition from this study). We preregistered Study 3 at https://aspredicted.org/31Q_TWK.

4.1 Method

4.1.1. Participants. Our final sample included 157 White adults (M_{age}=41.73 years, SD_{age}=14.20 years; 79 female, 77 male, 1 other) and 152 Black adults (M_{age}=37.07 years, SD_{age}=11.83 years; 73 female, 77 male, 2 other). We determined sample sizes prior to collecting data using G*Power, indicating 80% power for a two-tailed *t*-test and estimating an effect of Cohen's *d*=.50. Adults were recruited and completed the study via Prolific and received

compensation at a rate of $10 per hour. In addition to those described above, we tested and excluded 2 more adults for failing to answer attention check questions.[5]

4.1.2. Procedure. We randomly assigned participants to either an intentional or unintentional condition. We presented participants in the intentional condition with vignettes similar to those used in Studies 1 and 2, with the harm and intent underlying these actions made explicit. One example vignette read as follows:

> *A black couple came into a jewelry store and began looking at some of the jewelry on display. The lady behind the counter, who was white, then approached the couple and greeted them by asking, "Are you here to buy something today?" The black couple was upset by the lady's behavior. It turns out that the lady meant to have this effect on the couple. She truly did not want black people to shop at her store.*

Participants in the unintentional condition read similar vignettes that clarified that the harm was unintentional, as in the following example:

> *A black couple came into a jewelry store and began looking at some of the jewelry on display. The lady behind the counter, who was white, then approached the couple and greeted them by asking, "Are you here to buy something today?" The black couple was upset by the lady's behavior. It turns out that the lady did not mean to have this effect on the couple. She truly wanted to be a helpful salesperson.*

All the vignettes in the present study featured White individuals acting in intentionally or unintentionally harmful ways toward Black individuals. As Study 2 found a significant outcome for the difference placed on intent between Black and White participants (namely punishment judgments), participants completed the same punishment measure used in Study 2. Finally,

[5] Including all respondents in analyses yielded similar patterns as those reported in the main text.

participants answered a manipulation check question about the degree to which the vignettes they had read represented individuals who were trying to generate negative reactions intentionally (e.g., “In this study, you read about people who acted in ways that upset others. To what extent did those people mean to upset others on purpose?”) on a scale from 1=not at all to 7=a lot. Overall, participants in the intentional condition (*m*=5.16, *sd=2*.53) reported that the actions of the protagonists were more intentional than did participants in the unintentional condition (*m*=1.59, *sd*=1.34; *t*(232.34)=15.48, *p*=.001).

4.2 Results

To determine how participant race and the intentionality of the perpetrator's behavior shaped participants' punishment decisions, we averaged across the punishment items as described in Study 2 and entered these average scores into a 2 (Participants: White vs. Black) x 2 (Type of Harm: Intentional vs. Unintentional) ANOVA. Analyses revealed a main effect of Participant Race (*F*(1,309)=10.05, p=.001, *partial* η^2=.03): As in Study 2, White adults reported less punitiveness than did Black adults. We also observed a main effect of Type of Harm (*F*(1,309)=320.19, p=.001, *partial* η^2=.51): as in prior work (Cushman et al., 2013b; Payir & Heiphetz, 2022a), participants punished intentional harms more severely than unintentional harms. However, the interaction between Participant Race and Type of Harm did not reach significance (*F*(1,309)=.02, p=.88, *partial* η^2=.00; Fig. 5).

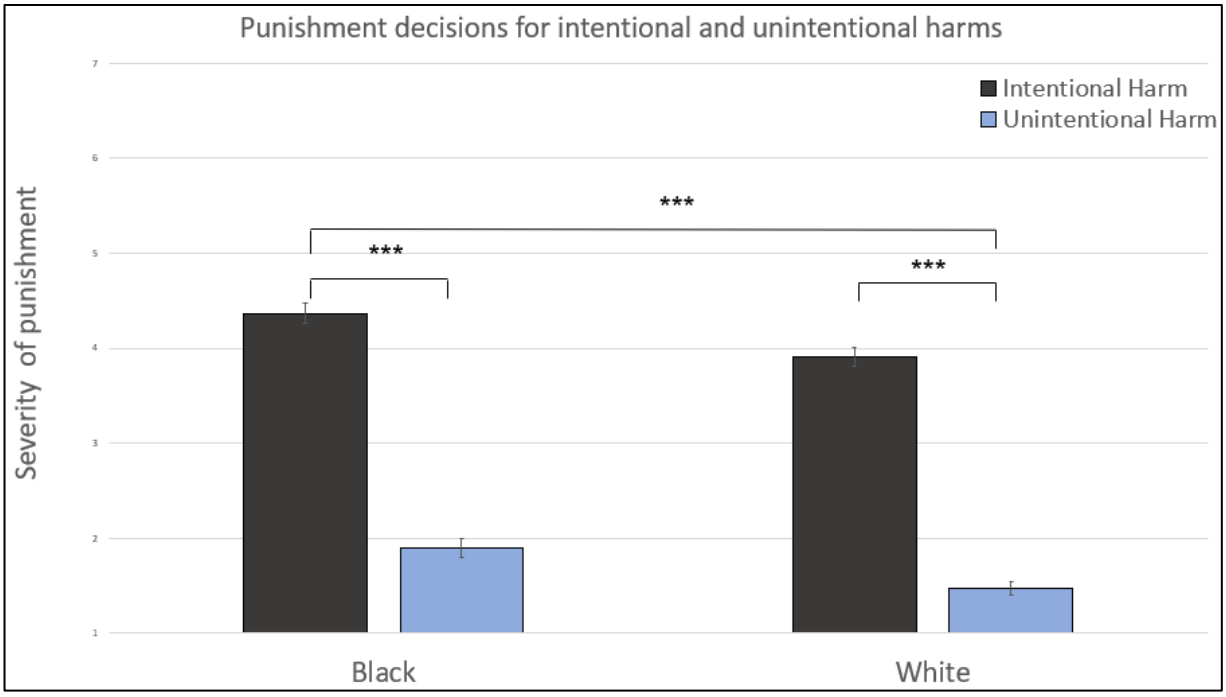

Fig 5. Means for Black and White participants' punishment decisions for intentional and unintentional intergroup harms, Study 3. Error bars represent 95% confidence intervals.

4.3 Discussion

Study 3 built on the previous studies by manipulating intent in vignettes of intergroup harms to investigate its causal role in shaping punishment judgments. When participants received explicit information about intent, they punished intentional intergroup harms significantly more severely than unintentional intergroup harms. This result expands scientific understanding of the role intent plays in moral cognition by demonstrating that intent causally shapes moral judgments in intergroup contexts, when explicit information about intent in is present. In this context information about intent also shapes punishment judgments. Study 4 extended Studies 1-3 by testing the generalizability of the results obtained in the intergroup condition and by probing a possible mechanism underlying participants' intent-based reasoning

Chapter 5: Study 4

Study 4 **(**preregistered at https://aspredicted.org/fqkz-qg96.pdf) built on Studies 1-3 in two ways. First, the intergroup conditions in Studies 1-3 only included White perpetrators and Black targets, a decision reflecting the fact that Black people commonly experience racist acts from White individuals (Desmond & Emirbayer, 2020; Gaertner & Dovidio, 2005; Norton & Sommers, 2011; Sue et al., 2019). While this design is high in ecological validity, it leaves open the question of whether results from these conditions would generalize to other intergroup contexts. Thus, Study 4 additionally asked participants about scenarios in which Black individuals harmed White individuals.

Second, Studies 1-3 examined Black and White adults' moral cognition and tested their judgments about perpetrator intent as a possible mechanism underlying this cognition. In contrast, Study 4 probed a mechanism hypothesized to underlie intent-based moral judgments. That is, Study 4 moved back a step from Studies 1-3, investigating why participants might reason about intent as they do.

Specifically, Study 4 asked whether threats to self-image might drive the emphasis that participants place on a perpetrator's intent. People typically perceive themselves as morally good (Heiphetz, 2019) and view moral goodness as central to their identity (Aquino & Reed, 2002; Heiphetz et al., 2017). When observing their ingroup members engaging in morally bad behaviors, people often experience threats to moral self-image (Chow et al., 2008; Knowles et al., 2014; Navarrete et al., 2004; Wohl et al., 2006). If an attempt to appease this threat underlies White participants' emphasis on a perpetrator's innocent intentions when observing ingroup members commit potentially racist acts, then providing an alternative route to threat appeasement (i.e., helping White participants feel like good people in some other way) may

reduce the extent to which they focus on perpetrator intent. Importantly, this element resembles what some scholars refer to as a form of "moral inoculation"—providing an alternative route to affirming moral identity that might preempt or mitigate the need to defensively emphasize intent. The study examines whether this process reduces reliance on intent-based reasoning, implying that participants' baseline focus on intent is, at least in part, a function of unresolved threat. In short, while the mechanism under investigation is a high-threat response, the intervention (moral affirmation) draws on the logic of inoculation to interrupt that threat response. The two processes are conceptually linked: inoculation functions as a protective buffer *because* the initial context is perceived as threatening to moral identity.

5.1 Method

5.1.1. Participants. Our final sample included 333 White adults (M_{age}=37.15 years, SD_{age}=12.46 years; 199 female, 130 male, 4 other) and 326 Black adults (M_{age}=38.52 years, SD_{age}=11.93 years; 150 female, 173 male, 1 other). We compensated participants at a rate of $10 per hour. We determined sample sizes prior to collecting data using G*Power, indicating 80% power for a two-tailed *t*-test and estimating an effect of Cohen's *d*=.50 but overrecruited in order to maintain power when accounting for exclusions. In addition to the data presented below, we excluded responses from 164 adults who failed to answer attention check questions from our analyses.[6]

5.1.2. Procedure. The procedure of Study 4 was identical to Study 1 except for the following differences. Before participants viewed the intergroup and intragroup harms, we randomly assigned half of the participants to a high threat condition wherein they were required to write 20 instances of morally good behaviors they performed over the past week. The

[6] Including all respondents in analyses yielded similar patterns as those reported in the main text.

remaining participants were assigned to a low threat condition wherein they were only required to provide 1 example of their own morally good behavior. Previous research corroborates this *ease of retrieval* paradigm as an effective tool for manipulating self-perceptions (Schwarz et al., 1991; Tybout et al., 2005). Specifically, thinking of 1 morally good behavior (the easier retrieval) should prompt individuals to believe they are more morally good, because for most people it will be easy to think of 1 morally good thing they have done. In contrast, thinking of 20 such items may be more difficult and may therefore lead participants to conclude that they are not particularly good, thus engendering moral self-image threat. After reporting 1 or 20 instances of prior morally good behavior, participants indicated the extent to which they felt like a good person; this item served as a manipulation check. Participants then responded to the moral judgment, punishment, and intent blocks as described in previous studies.

5.2 Results

As a manipulation check, we examined the degree to which participants believed they were a good person to determine whether or not the moral self-image threat was effective. However, participants in the low threat condition (m=8.34, sd=1.64) did not significantly differ from participants in the high threat condition (m=8.38, sd=1.54; t(663)=-.362, p=.718). To adhere with the pre-registration for this study, the results below examine this self-image threat variable; however, the general patterns reported below emerge even when this variable is removed (e.g., when it is not included in ANOVAs).

Like in previous studies, we averaged participants' responses to the moral judgment, punishment judgment, and intent emphasis questions and entered these averaged responses (separately) into a 2 (Participant Race: Black vs White) x 2 (Type of Harm: Black Perpetrator –

White Target vs White Perpetrator – Black Target) x 2 (Self-image Threat: Low vs High) mixed ANOVA.

For intent, analyses revealed a main effect of Participant Race ($F(1,651)=22.34$, $p<.001$, *partial* $\eta^{2=}.03$) and a significant interaction between Participant Race and Type of Harm ($F(1,651)=3.978$, $p=.047$, *partial* $\eta^{2=}.006$). No other main effects or interactions reached significance, ps>.065. To break down the Participant Race x Type of Harm interaction, we conducted two sets of tests. First, we investigated how White and, separately, Black participants reasoned about intent across both types of harm. White individuals prioritized intent more for White on Black harms ($m=5.24$, $sd=1.15$) than they did for Black on White harms ($m=5.01$, $sd=1.21$; $t(331)=4.564$, $p=.037$, Cohen's $d=.20$). This effect did not appear for Black individuals evaluating Black on White harms ($m=4.74$, $sd=1.18$) versus White on Black harms ($m=4.62$, $sd=1.40$; $t(324)=.867$, $p=.389$, Cohen's $d=.10$). Second, we investigated the extent to which Black versus White participants emphasized intent in the Black Perpetrator-White Target condition and, separately, in the White Perpetrator-Black Target condition. These analyses revealed that White participants emphasized intent somewhat more than Black participants in the Black Perpetrator-White Target condition ($t(325.55)=-1.978$ $p=.049$, Cohen's $d=.22$) and to a much greater degree in the White Perpetrator-Black Target condition ($t(309.69)=-4.366$, $p<.001$, Cohen's $d=.48$).

For moral judgments, analyses revealed a main effect of Type of Harm ($F(1,651)=96.75$, $p<.001$, *partial* $\eta^{2=}.13$): on average, participants judged actions where a White perpetrator harmed a Black target ($m=3.45$, $sd=.78$) to be less morally right than actions where a Black perpetrator harmed a White target ($m=4.04$, $sd=.72$; $t(659.87)=10.20$, $p<.001$, Cohen's $d=.79$). No other main effects or interactions reached significance, *p*s>.058.

For punishment judgments, analyses revealed three main effects (Fig. 6). First, we observed a main effect of Participant Race ($F(1,651)=102.02$, $p<.001$, *partial* $\eta^{2=}.13$): On average, Black participants reported that actions should be punished more severely than did White participants. Second, we observed a main effect of Type of Harm ($F(1,651)=30.23$, $p<.001$, *partial* $\eta^{2=}.04$): On average, participants reported that actions where a White perpetrator harmed a Black target should receive more severe punishment than actions where a Black perpetrator harmed a White target. Third, we observed a main effect of Moral Threat ($F(1,651)=96.75$, $p=.04$, *partial* $\eta^{2=}.005$): On average, participants in the high threat condition reported that behaviors should receive more severe punishment than participants in the low threat condition. No interactions reached significance, *p*s>.360.

In addition to investigating moral judgments and punishment decisions using the separate analyses presented above, the present study also correlated these variables to determine the degree to which moral cognition and decisions about morally relevant behaviors align with one another. Results showed a significant negative correlation between punishment judgments and moral judgments ($r=-0.394$, $p<.001$), indicating that less moral participants found a harm to be, the more severe were their punishment judgments.

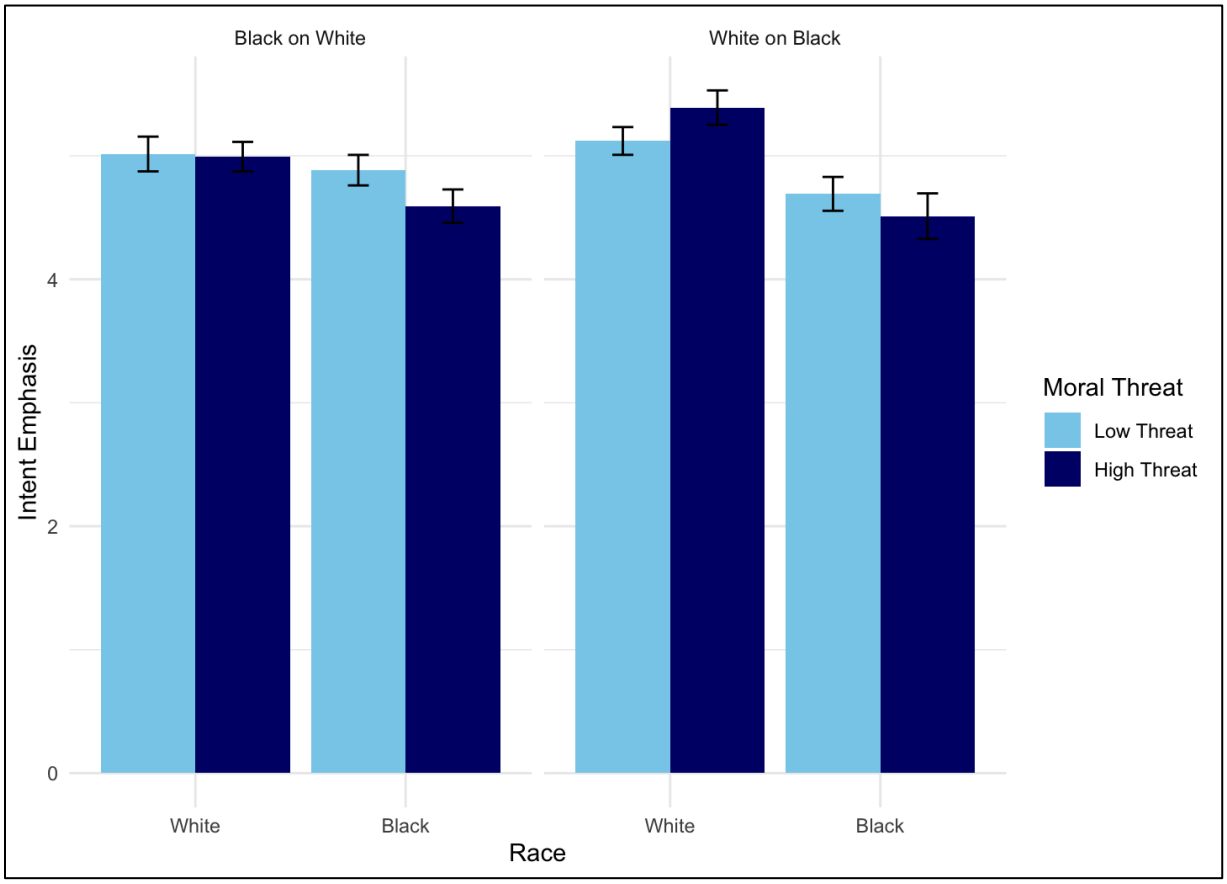

Fig 6. Means for participants' intent emphasis by type of intergroup harms, Study 4. Error bars represent 95% confidence intervals.

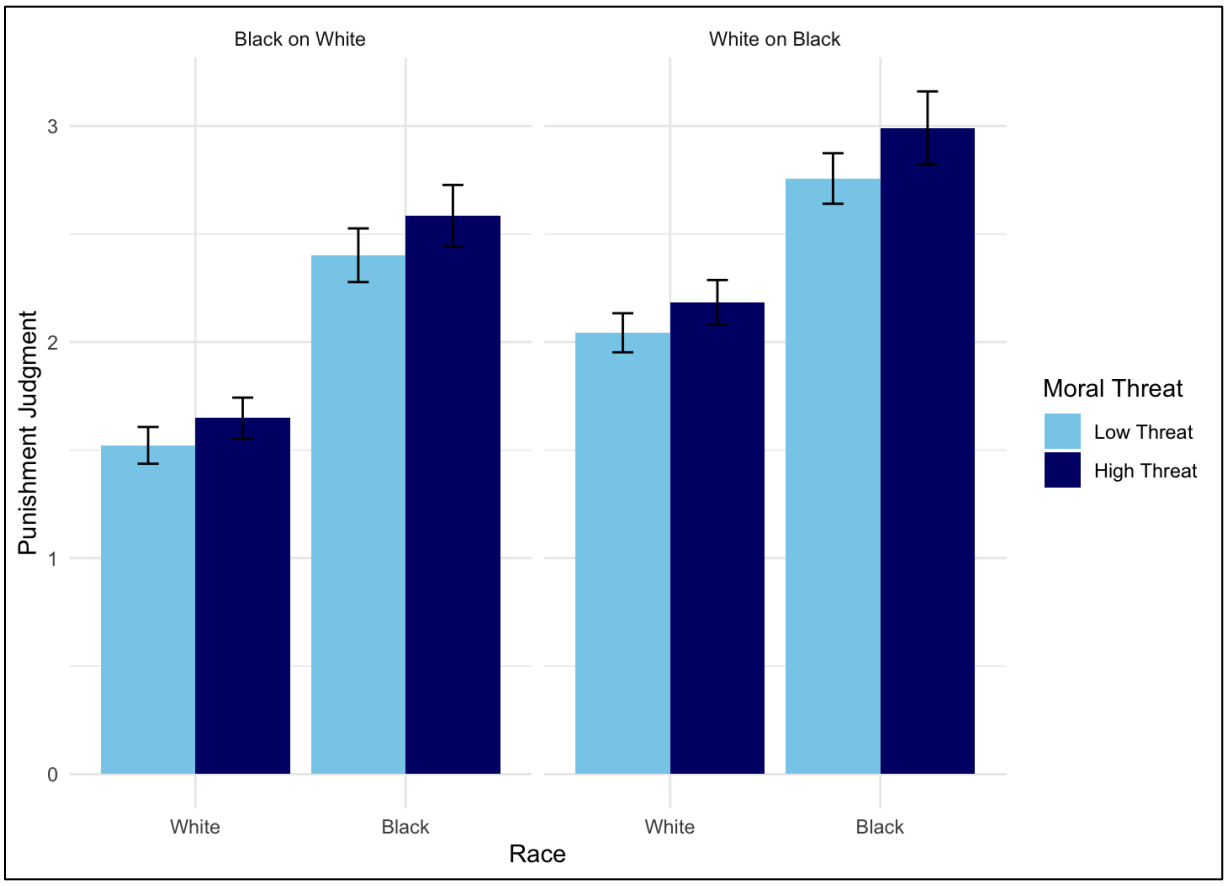

Fig 7. Means for severity of participants' punishment judgments by type of intergroup harms, Study 4. Error bars represent 95% confidence intervals.

5.3 Discussion

Study 4 expanded upon the findings of Studies 1-3 by directly investigating whether threats to moral self-image drive the differential emphasis on intent observed among White versus Black participants when evaluating intergroup harms. By attempting to induce moral self-image threat, we tested whether White participants' prioritization of intent could be reduced by buffering their moral self-perceptions. Contrary to our predictions, the moral self-image manipulation did not significantly alter participants' intent emphasis or moral judgments.

Importantly, the manipulation check did not yield significant results; thus, we cannot necessarily form strong conclusions about the relationship between participants' moral threat and intent emphasis. Nonetheless, several notable patterns emerged.

First, White participants consistently prioritized intent more than did Black participants, particularly when evaluating intergroup harms involving White perpetrators and Black victims. This aligns with the findings from Studies 1 and 2 and reinforces the idea that the mechanism underlying White participants' emphasis of intent may be distinct from a simple ingroup preference. If ingroup preferences drove White participants' responses, Black participants observing similar harms may have also shown similar ingroup preferences; however, such a finding did not emerge. Second, Black participants consistently punished intergroup harms more severely and judged them as more morally wrong than did White participants. This result may reflect the fact that Black individuals are often on the receiving end of these types of harms (Desmond & Emirbayer, 2020; Gaertner & Dovidio, 2005; Norton & Sommers, 2011) and may therefore think about and evaluate these harms differently from individuals who do not generally experience them (Freeman & Stewart, 2021; Molenberghs et al., 2016; Sue et al., 2019; Yudkin et al., 2016).

The lack of a significant effect for the moral self-image manipulation also invites further inquiry. It is possible that the retrieval task used to manipulate moral self-perceptions was insufficient to induce the desired self-image threat or that moral self-image plays a more nuanced role in shaping moral cognition in the context of intergroup harms than initially hypothesized. To further interrogate the role of moral self-image and the mechanisms underlying intent emphasis, Study 5 adopted a developmental lens. By focusing on the role that children's moral self-image may play in their intent-based judgments, Study 5 used a different methodological approach to

ask the same conceptual question as Study 4 (i.e., how moral self-image might be associated with intent-based reasoning). That is, while Study 4 used an experimental manipulation among adults, Study 5 relied on naturally occurring variation in moral self-image among children at different developmental timepoints to investigate how moral self-image might shape moral cognition.

Chapter 6: Study 5

Study 5 (preregistered at https://aspredicted.org/t5qq-7mf2.pdf) used developmental methods to further test the role of moral self-image in intent-based moral judgments. Specifically, this study focused on age-related differences in moral self-image to test whether White children's moral judgments in intergroup contexts became more sensitive to intent as the importance of a moral self-image increases. If that is the case, this finding would provide evidence that moral self-image threats underlie White participants' propensity to focus on perpetrator intent in intergroup contexts. If age-related differences do not emerge in Study 5, this finding would provide evidence against the hypothesis that moral self-image threats underlie participants' responses.

Moral self-image shifts between 4 and 9 years old. Indeed, some theories (e.g., Hardy & Carlo, 2011) argue that children on the younger end of this developmental window have a weaker moral identity and place less emphasis on being morally good than do children on the older side of this window. Several empirical results bolster this claim. For instance, in classic work on the "happy victimizer," Nunner-Winkler (2007) found that 4-year-olds expected transgressors to feel happy, whereas 8-year-olds expected more negative emotions. According to Nunner-Winkler's (2007) model, these results emerged because 8-year-olds themselves experience less positive emotions than do 4-year-olds after transgressing against others, an effect that shows that 8-year-olds place greater importance on moral behavior than do 4-year-olds. This interpretation is supported by empirical work showing that 8- to 9-year-olds are more likely than 5- to 6-year-olds to say that they would feel negative emotions after harming someone else (Keller et al., 2003). Additionally, preschoolers are less likely than children later in elementary

school to describe themselves in moral terms and view their own behaviors as morally relevant (Recchia et al., 2015; Sengsavang et al., 2015).

If White adults emphasize a perpetrator's innocent intent when evaluating intergroup transgressions because doing so buffers threats to moral self-image, then White preschoolers—who place less importance on being a "good person" than do older participants—should not show the same sensitivity to intent. In other words, being young may be akin to the manipulation in Study 4 that asked adults to recall a single instance of prior moral behavior to help them feel like good people. In both cases, participants are not expected to experience threats to moral self-image and may therefore see innocent intent as less important. If the data support this possibility, we should observe that older White participants (i.e., participants who are of an age to prioritize moral self-image and who also may experience threats to this self-image by virtue of reading about ingroup members behaving in ways that could appear racially biased) show more sensitivity to a perpetrator's innocent intent than younger White participants (who place less importance on being a good person and therefore do not need to take steps to buffer a threat to moral self-image) or any age group of Black participants (who are not expected to experience threats to their moral self-image after exposure to White people's potentially problematic behaviors). A different pattern of results would provide evidence against the idea that threats to moral self-image underlie the patterns of results observed in our earlier work. A developmental approach builds on the experimental approach in Study 4 by leveraging naturally occurring variations in moral self-image. If Study 4 did not show an effect of moral self-image threat because the manipulation we used did not adequately shift participants' moral self-image, relying on natural variation in moral self-image in Study 5 may yield different results (e.g., results

pointing toward moral self-image threat as a mechanism underlying intent-based moral reasoning in some contexts, if such a mechanism actually underlies participants' judgments about intent).

6.1 Method

6.1.1. Participants. Our final sample included 67 children between four and six years old (M_{age}= 5.22 years; SD_{age}=0.76; 30 female, 37 male) and 90 children between seven and nine years old (M_{age}=7.83 years; SD_{age}=0.86 years; 50 female, 40 male). We determined sample sizes prior to collecting data using G*Power, indicating 80% power for a 2 (Participant Age: Younger vs. Older) x 2 (Type of Harm: Intergroup vs. Intragroup) x 2 (Intent: Intentional vs. Unintentional) mixed design with repeated measures on the second and third factors estimating a medium effect of Cohen's f=.50.

6.1.2. Procedure. We first asked participants to complete a moral self-image measure taken from prior work (Sticker et al., 2021). Here, they learned about pairs of characters who were the same gender and approximate age as themselves. These characters were contrasted in their behavioral prosocial preferences such that one was always prosocial (e.g., indicated that they liked to share their crayons) and the other was always anti-social (e.g., indicated that they did not like to share their crayons). Participants viewed 9 such trials and, after each trial, indicated which character was more like them and the degree to which they believed this was the case.

Subsequently, we presented children with a series of 4 vignettes that described either an intergroup or an intragroup harm. These vignettes adapted the type of intergroup harms presented in previous studies to make them appropriate for a younger audience with visual supplements. The experimenter presented an intergroup (White on Black) or intragroup (White on White) harm with benign intent (e.g., a White character moved to a different part of the playground

when either a Black or White child sat next to them to play because they saw something fun to play with in a different part of the playground) or malicious intent (e.g., a White character moved to a different part of the playground because they did not want to sit next to a Black or White child). Participants then answered two questions about their punishment decisions (e.g., “Should [character] be punished for what they did?”, “Should [character] get into trouble for what they did?”). If they answered yes, participants indicated whether they felt this “a little”, “kind of”, or “a lot”. We coded responses on a scale from -3 (indicating that the punishment should be punished a lot) to 0 (indicating that the act should not be punished). The two moral judgment questions were structured similarly (e.g., “Was what the [child] did wrong or right?”, “Was what the [child] did bad or good”). We asked participants if they felt this 'a little', 'kind of', or 'very'. We coded these responses on a scale from -2.5 (indicating that the act was very wrong/bad) to +2.5 (indicating that the act was very right/good). This coding scheme allowed for the distance between responses to be consistent. For instance, the difference between -0.5, indicating that the participant believed that the behavior was "a little" negative, and +0.5, indicating that the participant believed that the behavior was "a little" positive, was the same as the distance between any two scale points that were on the same side of zero as each other.

6.2 Results

As in previous studies, we averaged participants' responses to the moral judgment items and, separately, the punishment items and entered these averaged responses (separately) into a 2 (Participant Age: Younger vs Older) x 2 (Type of Harm: Intergroup vs Intragroup) x 2 (Intent: Intentional vs Unintentional) mixed ANOVA with repeated measures on the second and third factors.

For moral judgment, analyses revealed a main effect of Intent ($F(1,122)=57.56$, $p<.001$, *partial* $\eta^{2=}.32$): on average, participants rated intentional harms as morally worse than unintentional harms. This main effect was qualified by a Participant Age x Intent interaction ($F(1,122)=10.05$, $p=.002$, partial $\eta^2=.014$), which, in turn, was qualified by a Participant Age x Intent x Type of Harm interaction ($F(1,122)=7.861$, $p=.006$, partial $\eta^2=.010$). No other main effects or interactions reached significance, ps>.478.

To break down the three-way interaction, we examined the effects of Intent and Type of Harm across different age groups. This analysis revealed that older children rated intentional intergroup harms (m=-1.79, sd=.76) as more morally wrong than did younger participants (m=-0.83, sd=1.68; t(13-)=4.15, $p< .001$, Cohen's d=0.19). However, this effect was not present in older children (m=-1.16, sd=1.20) when rating intentional intragroup harms as compared to younger children (m=-.98, sd=1.60; t(149)=.87, p=.386, Cohen's d=0.14). Similarly, significant differences between older and younger participants did not emerge when participants rated the moral wrongness of unintentional intergroup harms (older participants: m=-.21, sd=1.47; younger participants: m=-.41, sd=1.51; t(153)=-.82, p=.411, Cohen's d=.16) or unintentional intragroup harms (older participants: m=-.44, sd=1.51; younger participants: m=-.46, sd=1.66; t(149)=-.09, p=.926). In other words, older participants—i.e., participants who are particularly likely to emphasize the importance of moral self-image (Hardy & Carlo, 2011)—showed more sensitivity to intent than did younger participants when evaluating intergroup harms. This result did not emerge in evaluations of intragroup harms, suggesting that the age-related difference may not arise from age-related changes in intent-based reasoning broadly construed; if this were the case, older children would have shown more sensitivity to intent regardless of whether they evaluated intergroup or intragroup harm.

For punishment judgment, analyses revealed a significant main effect of Intent ($F(1,123)=24.043$, $p<.001$, partial $\eta^2=.16$): on average, participants reported that intentional harms should receive more severe punishment than unintentional harms. Analyses also revealed a main effect of Type of Harm ($F(1,123)=8.34$, $p=.005$, partial $\eta^2=.06$): on average, participants reported that intergroup harms should receive more severe punishment than intragroup harms. Neither the main effect of Participant Age nor any two-way interactions reached significance, ps>.090.

The significant main effects of Intent and Type of Harm were qualified by a Participant Age x Type of Harm x Intent interaction ($F(1,123)=5.60$, $p=.019$, partial $\eta^2=.007$, Fig. 7). To break down the three-way interaction, we examined the effects of Intent and Type of Harm across different age groups. The only difference that emerged between older versus younger children's punishment decisions was in regard to unintentional intergroup harms; here, older participants ($m=-.31$, $sd=.63$) provided less harsh punishment decisions than did younger participants ($m=-.63$, $sd=.88$; $t(123)=-2.73$, $p=.007$, Cohen's $d=.44$). While the results reported above indicated that older participants rated intentional intergroup harms as more morally wrong than did younger participants, the results regarding punishment decisions did not show this result; older and younger participants did not differ in the extent to which they reported that the perpetrators of intentional intergroup harms should receive punishment (older participants: $m=-.95$, $sd=1.08$; younger participants: $m=-.84$, $sd=1.01$; $t(130)=.588$, $p=.557$, Cohen's $d=.10$). Older and younger participants also did not differ in their punishment decisions regarding intentional intragroup harms (older: $m=-.60$, $sd=.90$; younger: $m=-.82$, $sd=1.02$; $t(150)=-1.43$, $p=.154$, Cohen's $d=.23$) or in their punishment decisions regarding unintentional intragroup harms (older: $m=-.38$, $sd=.69$; $m=-.55$, $sd=.84$; $t(150)=-1.34$, $p=.181$, Cohen's $d=.22$). We also tested the effect

of intent and type of harm within each age group. Older children punished intentional intergroup harms (*m*=-.95, *sd*=.1.06) significantly more severely than intentional intragroup harms (*m*=-.51, *sd*=.89; *t*(153)=2.31, *p*=.02, Cohen's *d*=.14). This difference was not present in older children when comparing unintentional intergroup harms (*m*=-.30, *sd*=.64) and unintentional intragroup harms (*m*=-.31, *sd*=.68; *t*(171.55)=.74, *p*=.463, Cohen's *d*=.04). No significant differences appeared within younger children (*p*s >.557).

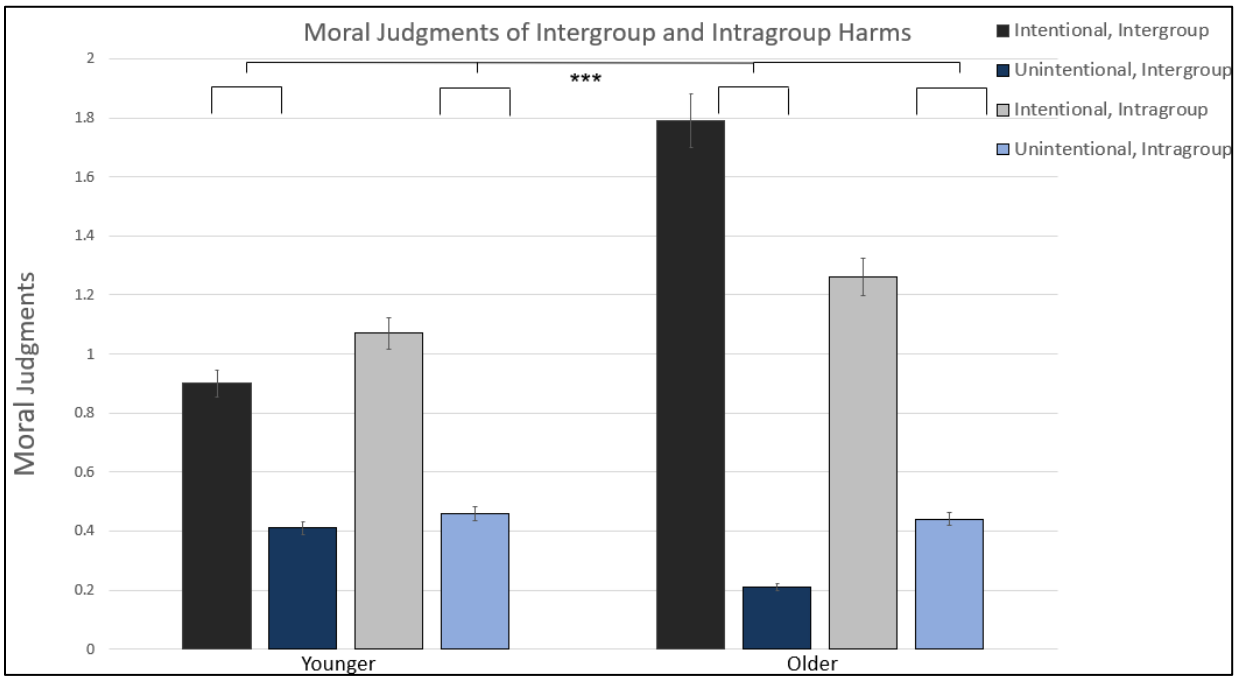

Fig 7. Means for participants' moral beliefs by type of intergroup harms, age group, and intent, Study 5. Error bars represent 95% confidence intervals.

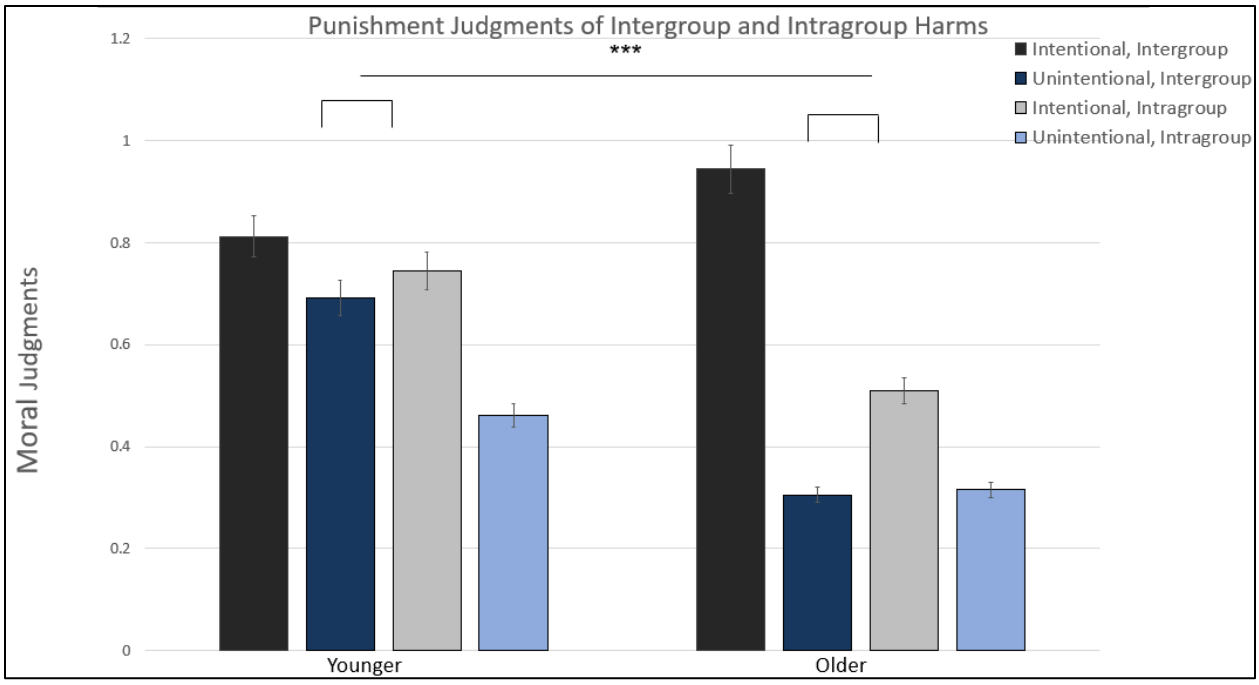

Fig 8. Means for participants' punishment beliefs by type of intergroup harms, age group, and intent, Study 5. Error bars represent 95% confidence intervals.

The present study also correlated moral and punishment judgments to determine the degree to which these beliefs aligned when evaluating intergroup and intragroup harms. A significant positive correlation emerged between punishment judgments and moral judgments when collapsing across the intent and type of harm conditions (r=.492, p<.001), indicating that the more immoral participants found a harm to be, the more severe were their punishment judgments, on average. While moral judgments are often influenced by the perceived intent behind an action, punishment decisions may incorporate additional considerations, such as the severity of the harm, the context of the behavior, and the perceived need for deterrence or retribution. Older children in Study 5 rated intentional intergroup harms as more morally wrong than younger children, reflecting their greater sensitivity to intent. However, when it came to

punishment, older children did not consistently impose harsher penalties for intentional intergroup harms compared to younger children. This discrepancy suggests that while older children may recognize the moral wrongness of intentional harms, they may also consider mitigating factors, such as the potential for rehabilitation or the broader social implications of punishment, when deciding on appropriate consequences. Additionally, the lack of significant differences in punishment for unintentional harms across age groups may indicate that children, regardless of age, are less inclined to punish actions they perceive as accidental, even if they recognize the moral ambiguity of such behaviors.

Finally, we conducted a mediation analysis with 5,000 bootstrapped samples. This mediation included age as the predictor variable, children's moral self-image as the mediator, and moral judgments of unintentional intergroup harms as the outcome variable, as this condition best encapsulates the prototypical context of the harm (intergroup with a potentially benign justification) and drives our main question as to the mechanistic role of moral self-image for the emphasis of intent in intergroup harms. This mediation model did not reveal significant indirect effect (*B*=-.007, p=.733, 95% CI[-.048,.034]). This was also true of children's punishment judgments; moral self-image did not significantly mediate the relationship between intent emphasis and punishment (*B*=-.002, p=.867, 95% CI[-.022,.018]).[7] Additionally, comparing younger children (*m*=.65, *sd*=.67) to older children (*m*=.69, *sd*=.22) in a t-test did not reveal an overall significant difference with regards to moral self-image (*t*(155)=-1.01, *p*=.313).

[7] We replicated these mediation analyses with the other substantive conditions in Study 5. The subsequent models were not significant for either the moral judgment (*p*s>.282) or punishment judgment (*p*s>.412) outcome variables.

6.3 Discussion

Study 5 extended the findings of prior studies by investigating how potential age-related difference in moral self-image might shape moral judgments and punishment decisions regarding intentional and unintentional harms performed in intergroup and intragroup contexts. The results revealed several notable patterns. First, as in prior work (e.g., Baird & Astington, 2004; Killen et al., 2011; Payir & Heiphetz, 2022), participants judged intentional harms more harshly than unintentional harms. Second, the interaction between Intent, Type of Harm, and Participant Age revealed that older children resembled believed intergroup harms as more morally wrong than intragroup harms and more worthy of punishment. Interestingly, this stands in contrast to the pattern of results found in Studies 1 and 2 wherein while adults believed ambiguously intentional intergroup harms to be more morally wrong than ambiguously intentional intragroup harms, they nonetheless punished these harms less severely.

Of note, older children's punishment of intergroup harms is more severe when intentional, but did not significantly differ from their punishment of intragroup harms when unintentional. This may indicate that children undergo a shift between this period in development and adulthood such that they find potentially unintentional intergroup harms less worthy of severe punishment as compared to intragroup harms.

General Discussion

The present work investigated how individuals think about intent when making morally relevant judgments about harms in intergroup and intragroup contexts. Study 1 found that White adults prioritized intent to a greater degree than did Black adults with regards to intergroup harms and further showed that intent emphasis mediated the relationship between type of harm (intergroup versus intragroup) and moral judgment. Study 2 found that Black and White individuals punished intergroup harms less severely than intragroup harms and that White participants punished harms overall less severely than did Black participants. Study 3 manipulated information about intent and found a significant of effect of both race and type of harm, such that both Black and White participants punished intentional intergroup harms more severely than unintentional intergroup harms and White participants punished intentional intergroup harms less severely than did Black participants. Furthermore, intent did not mediate the relationship between type of harm and punishment judgments for either White or Black participants. Study 4 revealed that White participants emphasized intent more when evaluating harms involving their ingroup as perpetrators and this tendency remained constant in contexts where they were prompted to make both moral and punishment judgments. However, manipulating moral self-image did not significantly influence results. Finally, Study 5 complemented these findings by testing moral cognition across development. Older children differed from young children in that they believed that intergroup harms were more wrong than intragroup harms and they were also more likely on average to punish them more severely, where young children did not differentiate between unintentional and intentional harms. However, the absence of mediation effects for moral self-image across age groups, in conjunction with the null results regarding moral self-image in Study 4, suggests that other

factors may play a stronger role in underlying the development of moral judgment. Taken together, these studies highlight the role of group membership in moral cognition and emphasize the importance of integrating moral psychology with the social science of intergroup bias.

Implications for Moral Psychology

The present studies provide a differing perspective on concepts commonly held in moral psychology. While much of the existing literature emphasizes the centrality of intent in moral judgments, particularly among adults and older children (e.g., Cushman, 2008; Cushman et al., 2013; Killen et al., 2011), our findings reveal a more nuanced picture. Intent does play a role in moral reasoning, but its influence is context-dependent and varies across individuals and types of harm.

In Study 1, intent mediated the relationship between type of harm (intergroup versus intragroup) and moral judgments, but only for White participants. This suggests that intent matters more for White individuals when evaluating intergroup harms as compared to intragroup harms, and this preference is distinct from ingroup preference as it was not present in Black individuals evaluating harms where Black perpetrators harmed White victims. This result is consistent with work showing that individuals who experience interpersonal transgressions highlight different aspects of those experiences than do perpetrators (e.g., victims tend to portray the incident as part of a broader context, whereas perpetrators tend to portray the incident as isolated; (Baumeister et al., 1990) and extends results highlighting differences in perspective to an intergroup context. Study 2 further highlighted the role of intent, showing that both Black and White participants punished intentional intergroup harms more severely than intragroup harms. However, White participants punished intentional intergroup harms less severely than Black participants, suggesting that intent may serve as a mitigating factor for White individuals but not

for Black individuals. While not directly examining the lived experience of participants in naturalistic settings, the observed divergence illustrates how group membership might shape the weight individuals assign to intent in situations that are common to the lived experience of minoritized groups in the United States. Interestingly, in Studies 2 and 4, intent did not mediate the relationship between type of harm and punishment judgments for either group. Aligning with previous research (Frisch et al., 2021),this result suggests that intent matters less when the focus shifts from moral wrongness to punishment severity. Participants may judge intergroup harms as more morally wrong due to their broader social implications, but they may be less willing to punish these harms severely, especially when intent is ambiguous and the punishment may fall on ingroup members. This discrepancy aligns with prior work showing that distinct psychological processes drive judgments of wrongness versus punishment (Kneer & Machery, 2017; Kneer & Skoczen, 2023).

Taken together, these findings indicate that intent matters especially when individuals are evaluating the moral wrongness of harms, particularly in intergroup contexts where the social implications of an intergroup harm are greater than those of an intragroup harm. Previous studies in moral psychology focusing on perceptions of intent and on moral judgments more generally have commonly focused on the interpersonal level (Cushman et al., 2013b; Killen et al., 2011; Payir & Heiphetz, 2022). However, many harms do occur along group membership lines (Desmond & Emirbayer, 2020; Gaertner & Dovidio, 2005; Norton & Sommers, 2011), and insight into how people judge within-group harm may not extrapolate to intergroup harm. For instance, individuals may respond more leniently to harms committed by ingroup members (Okonofua & Eberhardt, 2015; Sidanius & Pratto, 1999; Yudkin et al., 2016), particularly when they can make benign attributions for those harms (e.g., that the harmful behavior was

unintentional). The present studies add to this body of work by directly examining how members of different groups morally evaluate instances of intergroup and intragroup harm. One possible implication is that in contexts where individuals make moral evaluations, they attune not only to information directly related to the harm observed, but to information about the group memberships of the individuals involved. Thus, moral psychology can improve its understanding of morally relevant behavior by incorporating perspectives from intergroup research on how individuals make judgments in intergroup and intragroup contexts. Doing so enables moral psychology to better understand how individuals evaluate intergroup harms differently from intragroup harms across situations that they might reasonably experience in everyday life (e.g., while shopping or riding public transportation, as reflected in the stimuli used in the current work).

Implications for Intergroup Relations

The present work also possesses clear implications for intergroup processes. While explicit endorsements of negative beliefs about racial or ethnic groups in the U.S. declined between the 1980s and the mid-2010s, a more covert and ambiguous form of racism was becoming more prevalent in that span (Bobo et al., 2012; Bonilla-Silva, 2006; Joseph, 2017; Nadal et al., 2020; Testa, 2023). This subtle shift in presentation means that members of minoritized groups often experience instances of racism and discrimination that go unnoticed or unacknowledged by members of dominant groups (Gaertner & Dovidio, 2005). These instances of ambiguous racism pose unique risks to members of minoritized groups due to their ubiquity and can undermine the ability of members of targeted groups to appeal to both social and institutional mechanisms for holding perpetrators accountable (Sue et al., 2007). For example, systemic racism in the criminal legal system often operates through seemingly race-neutral

policies, such as mandatory minimum sentencing for drug offenses (Alexander, 2012). While these policies appear impartial, they disproportionately target Black and Brown communities, leading to mass incarceration and the disenfranchisement of millions. However, because the racism embedded in these policies is subtle and structural, it becomes difficult for affected individuals to challenge or hold institutions accountable. The ambiguity of such harms—coupled with the lack of explicit intent—allows systemic injustices to persist unchecked, leaving minoritized groups without clear avenues for redress. This dynamic mirrors the challenges faced in everyday interpersonal interactions, where ambiguous racist behaviors (e.g., microaggressions) are often dismissed or excused, further marginalizing those who experience them

While moral psychology often emphasizes intent in various moral reasoning processes (Cushman et al., 2013a; Payir & Heiphetz, 2022; Young & Saxe, 2011), the degree to which intent *should* be emphasized in matters of intergroup harms remains unclear, especially when thinking about these more covert types of harms. One specific issue with prioritizing intent in this manner is that it centers transgressors rather than their targets. This dynamic is also present within intergroup research on microaggressions and other related intergroup harms. When the meaning of a concept is derived from the harmful actions of the one who commits an egregious act, then the agent committing the action is the focus (Freeman & Stewart, 2021; Sue et al., 2007; Sue & Spanierman, 2020). Consequently, the recipient becomes a secondary subject of interest. If the fields of intergroup research and moral psychology are interested in speaking to the experience of groups that experience ambiguous intergroup harms like people of color, people of color, not the perpetrators of intergroup harm, need to be the focal point of the discussion.

The present work found that on average, Black and White individuals punish intergroup harms less severely than intragroup harms that occur in similarly ambiguous contexts. Work in the field of intergroup relations does not integrate from moral psychology when trying to explain why individuals respond in certain ways in intergroup contexts. The present work attempted to bridge this theoretical gap by proffering threat to moral self-image (and subsequent compensatory behavior) as a mechanism underlying differences on morally relevant judgments that fall along group membership lines. Testing moral self-image as a factor provides insight into the degree to which individuals' moral judgments reflect the desire to maintain a positive moral self-image in intergroup contexts**.** This distinction may also clarify instances where people accept benign intent as a justification or excuse for harms committed by their ingroup members as well as instances where they might not (and why individuals that are not members of those ingroups might be equally less likely to accept explanations from outgroup members). By identifying how moral self-image shapes these asymmetries, future research can inform strategies for promoting more equitable moral and punishment judgments in intergroup settings.

Implications for Developmental Psychology

Our studies provide critical insights into how moral judgments about intergroup and intragroup harms develop across childhood and into adulthood. Prior research has established that young children rely heavily on outcome-based reasoning, with a gradual shift toward intent-based moral judgments beginning around 4-5 years of age (Cushman et al., 2013; Killen et al., 2011). However, our findings suggest that this developmental trajectory may not be uniform across intergroup and intragroup contexts. Study 5 revealed that older children, unlike adults, judged intergroup harms to be more morally wrong than intragroup harms and were more likely to impose severe punishments for intergroup transgressions versus intragroup transgressions.

This pattern stands in contrast to adults, who tended to punish intergroup harms less severely than intragroup harms. These findings suggest that children may be more attuned to the perceived injustice of intergroup harms and less likely to mitigate punishment based on intent. This result aligns with research indicating that children are often more egalitarian in their moral reasoning before gradually adopting the biases and justifications present in adult social environments (Rutland et al., 2010; Skinner & Meltzoff, 2019).

One possible explanation for this developmental shift is that as children grow, they become increasingly socialized into norms that emphasize intent in moral reasoning, particularly in intergroup contexts. Given that White adults in our studies placed greater emphasis on intent when evaluating intergroup harms versus intragroup harms, it is likely that these norms reflect broader societal tendencies to justify or downplay ingroup wrongdoing in order to protect the perception of one's group. The observed developmental differences raise important questions about when and how children begin to adopt these patterns of moral judgments and the extent to which exposure to societal narratives about intergroup relations plays a role in shaping these shifts. Future research should investigate how children from different racial and ethnic backgrounds internalize messages about intent and harm, particularly in relation to historical and contemporary intergroup inequalities.

These findings also have educational implications. Interventions designed to promote moral reasoning in children often emphasize the role of intent in moral evaluations (e.g., emphasizing that accidents are less blameworthy than intentional harms; Arsenio & Lover, 1995; Schonert-Reichl et al., 2012). However, in the context of intergroup relations, an overemphasis on intent may inadvertently reinforce biases that excuse or minimize intergroup harm. Educational programs should consider incorporating discussions about the broader social

consequences of harm, helping children understand that even unintentional actions can perpetuate systemic inequalities. Encouraging children to focus on the perspectives of those harmed rather than solely on the motivations of the perpetrator may foster a more nuanced and socially conscious approach to moral reasoning.

Additionally, research should examine how common interventions—such as intergroup contact, exposure to diverse perspectives, and cultural narratives—influence the role of intent in the evaluation of intergroup and intragroup harms. Currently, intergroup contact interventions often focus on reducing prejudice by fostering positive interactions between members of different groups (e.g., Pettigrew & Tropp, 2006). However, based on the findings of this work, such interventions should also explicitly address how individuals evaluate intent in intergroup contexts. For example, programs could include activities that encourage participants to reflect on how intent is perceived differently across groups and how unintentional harms can still perpetuate inequality. Similarly, exposure to diverse perspectives is commonly used to broaden individuals' understanding of others' experiences (e.g., Todd & Galinsky, 2014). Yet, this work suggests that such interventions should go further by teaching participants to critically examine how intent is weighed in moral judgments, particularly in cases where harms are ambiguous or unintentional. Finally, cultural narratives often shape societal norms around morality and intent (e.g., Hammack, 2011). Often prominent narratives frequently emphasize the importance of intent in excusing or justifying harmful behaviors, but this work highlights the need to reframe these narratives to prioritize the perspectives of those harmed, especially in intergroup contexts. For instance, media and educational programs could highlight stories that illustrate the impact of unintentional harms on marginalized groups, fostering a more nuanced understanding of intent and accountability. By integrating these insights, interventions can better address the

complexities of moral reasoning in intergroup settings and promote more equitable evaluations of harm.

Limitations and Future Directions

The present studies provide novel insight into Black and White adults', and White children's, moral judgments. In so doing, the current work strengthens the bridges between moral psychology and work on intergroup bias. Additionally, the progression of studies addressed some limitations. For instance, Studies 1-2 asked participants how much emphasis they placed on intent. One benefit of this approach is that it allowed us to measure naturally occurring variation in intent-based judgments and determine how this variation might relate to moral judgments and punishment decisions. However, one drawback of this strategy is that a correlational approach does not allow for causal claims. Therefore, Studies 3 and 5 manipulated information about intent. This approach did not allow for insight into the extent to which Black versus White participants prioritize intent, but it did allow for causal interpretations.

Studies 4-5 aimed to address some remaining limitations in Studies 1-3. For instance, Studies 1-3 investigated a focus on intent as a mechanism underlying Black and White participants' responses to intergroup and intragroup harms. However, those studies did not ask why Black and White participants might reason about intent in the ways that they do. As such, Studies 4-5 probed a mechanism—namely, threats to moral self-image—that may underlie intent-based judgments. However, the ease-of-retrieval paradigm used in Study 4 was not able to successfully manipulate participants' moral self-image. Future work could improve on this by employing more robust manipulations of moral self-image in adults, such as direct feedback tasks (e.g., providing participants with false feedback about their moral character based on a bogus personality test; Jordan et al., 2015) or moral identity priming (e.g., having participants

write about a time they acted immorally or morally, depending on the condition; Sachdeva et al., 2009). For children, developmentally appropriate methods could include storytelling interventions that highlight moral self-concept (e.g., asking children to reflect on their own moral behaviors in hypothetical scenarios; Sticker et al., 2021) or observational tasks that measure children's reactions to moral transgressions in group settings (Recchia et al., 2015). These approaches would allow researchers to more effectively test the role of moral self-image in shaping intent-based judgments in both developmental and intergroup contexts.

One limitation of the earlier studies is that we tested only one form of intergroup harm—where a White perpetrator harmed a Black target. This design captured a socially relevant dynamic, but it did not allow us to determine whether the observed effects were specific to this particular harm configuration or whether they generalized to cases where a Black perpetrator harmed a White target. By incorporating this additional condition in later Study 4, we were able to examine whether participants' moral judgments were shaped by group membership asymmetrically or if similar patterns emerged across different intergroup harm contexts. This extension provided critical insight into whether individuals apply the same moral reasoning principles across intergroup transgressions or whether responses are influenced by broader sociocultural narratives about race and harm.

While moral self-image provides a useful framework for understanding why people might emphasize intent more in instances of harm where ingroup members represent perpetrators, it was not a particularly strong mechanism underlying Study 5's results, and other mechanisms may contribute more to the observed differences in moral judgments. For instance, social identity threat—the fear of one's group being negatively evaluated (Steele et al., 2002; Branscombe et al., 1999)—may play a key role, particularly for White participants, who may be

motivated to downplay the severity of intergroup harms to protect their ingroup's broader social reputation. This motivation is distinct from moral identity threat which arises from a fear that one's personal or collective moral integrity is being called into question (Ellemers et al., 2013; Monin & Jordan, 2009)—and may drive White participants to emphasize intent as a means of preserving a positive moral conceptualization of their group. Similarly, system justification theory (Jost & Banaji, 1994) suggests that individuals may engage in motivated reasoning to uphold existing social hierarchies, which could explain why White participants prioritize intent more in intergroup contexts—potentially as a way to rationalize and minimize the harm done by their ingroup. Additionally, perceptions of group-based victimhood (Sullivan et al., 2012) may shape responses among both Black and White participants, with Black participants viewing intergroup harms as part of a broader pattern of racial injustice, while White participants may be more likely to interpret these harms as isolated incidents. Future research should investigate how these mechanisms interact with moral self-image and the degree to which intergroup mechanisms processes distinct from the specific context of a given harm drive moral evaluations in intergroup and intragroup contexts.

Finally, the present studies examined moral cognition among Black and White participants with regard to harms that occur without explicit intent. By investigating how Black and White individuals navigate moral evaluations of ambiguous harms, these studies provide insight into whether and how intergroup dynamics shape moral reasoning beyond explicit transgressions. The benefit of doing so is that it leverages the lived experience of Black individuals, namely that the intergroup harms they experience on a frequent basis (Desmond & Emirbayer, 2020; Gaertner & Dovidio, 2005; Norton & Sommers, 2011; Sue et al., 2019) are often marked by an ambiguous intent. It is often unclear whether these harms are unintentional or

reveal a harbored prejudice. However, Black and White adults do not comprise the entirety of the United States, and future work can probe how other group memberships may shape moral cognition. For instance, research should explore how Latino, Indigenous, and Asian American individuals navigate moral judgments in intergroup contexts, given their distinct histories of racialization and experiences with systemic inequality. Latino individuals, for example, may show patterns similar to Black individuals due to shared experiences of discrimination and racial marginalization, whereas Indigenous individuals—who have historically experienced relatively more erasure (Fryberg et al., 2024)—may exhibit even stronger sensitivity to the moral implications of ambiguous intergroup harms. Additionally, Asian American individuals, who are often racialized through the lens of the "model minority" stereotype (Park et al., 2021; Russel & McCurdy, 2023), may demonstrate an emphasis on intent in intergroup harm contexts for both White and Asian perpetrator contexts as compared to Black and Latino populations, particularly in contexts where they perpetrate intergroup harm against Black or Latino populations versus experience harm from White populations. Investigating these groups would allow for a more comprehensive understanding of how moral cognition is shaped by the lived experience of different racial and ethnic communities.

Conclusion

The current studies provide valuable insights into the complexities of moral judgment, intent, and intergroup processes. Black and White adults evaluate intergroup and intragroup harms differently with regard to how they prioritize intent (such that White individuals prioritize intent more for intergroup harms than intragroup harms when making moral judgments, whereas Black individuals do not differentiate) and whether they are making punishment versus moral judgments (where White individuals punish intergroup harms less severely than Black

individuals). Additionally, these studies reveal a developmental difference such that older children differentiated themselves from younger children (and the adults in Study 1) in their more severe moral judgments of intergroup versus intragroup harms. Understanding how children develop intent-based reasoning in intergroup settings and how socialization processes shape these trajectories will be critical for uncovering the roots of moral biases. These findings underscore the need to incorporate intergroup perspectives into scientific understanding of moral reasoning to reflect the experiences of individuals who regularly face intergroup harms. Additionally, centering intent as the critical factor in judgments of intergroup harm poses the risk of decentering marginalized individuals in favor of transgressors in instances where they face harm. Moreover, a developmental approach is essential for identifying how and when these patterns emerge, how they change over the course of development, and potentially informing interventions that could mitigate biases before they become deeply ingrained. Taken together, this work indicates that moral cognition in intergroup contexts cannot be fully understood through a single theoretical lens—whether focusing solely on intent, group membership, or developmental processes. Instead, it requires an integrated approach that considers how intent, motivational considerations like self-image, and intergroup processes interact across development, particularly in shaping how individuals evaluate and respond to harms that reinforce or challenge social hierarchies.

References

Adler, A. (2006). Fundamentals of individual psychology. In S. Slavik & J. Carlson (Eds.), *Readings in the Theory of Individual Psychology* (pp. 33–43). Routledge/Taylor & Francis Group.

Ames, D. L., & Fiske, S. T. (2013). Intentional Harms Are Worse, Even When They're Not. *Psychological Science*, *24*(9), 1755–1762. https://doi.org/10.1177/0956797613480507

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Arsenio, W., & Lover, A. (1995). Children's conceptions of sociomoral affect: Happy victimizers, mixed emotions, and other expectancies. *Morality in Everyday Life: Developmental Perspectives.*, 87–128.

Barkan, R., Ayal, S., Gino, F., & Ariely, D. (2012). The pot calling the kettle black: Distancing response to ethical dissonance. *Journal of Experimental Psychology: General*, *141*(4), 757–773. https://doi.org/10.1037/a0027588

Baumeister, R. F., Stillwell, A., & Wotman, S. R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *Journal of Personality and Social Psychology*, *59*(5), 994–1005. https://doi.org/10.1037/0022-3514.59.5.994

Bergman, R. (2002). Why Be Moral? A Conceptual Model from Developmental Psychology. *Human Development*, *45*(2), 104–124. https://doi.org/10.1159/000048157

Bobo, L. D., Charles, C. Z., Krysan, M., Simmons, A. D., & Fredrickson, G. M. (2012). The real record on racial attitudes. *Social Trends in American Life: Findings from the General Social Survey since 1972*, 38–83.

Bonilla-Silva, E. (2006). *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States*. Rowman & Littlefield Publishers. https://books.google.com/books?id=HKJE4rVZG1EC

Čehajić-Clancy, S., & Olsson, A. (2024). Threaten and affirm: The role of ingroup moral exemplars for promoting prosocial intergroup behavior through affirming moral identity. *Group Processes & Intergroup Relations*, *27*(1), 99–117. https://doi.org/10.1177/13684302221148397

Chow, R. M., Lowery, B. S., & Knowles, E. D. (2008). The two faces of dominance: The differential effect of ingroup superiority and outgroup inferiority on dominant-group identity and group esteem. *Journal of Experimental Social Psychology*, *44*(4), 1073–1081. https://doi.org/10.1016/j.jesp.2007.11.002

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013a). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013b). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008

Desmond, M., & Emirbayer, M. (2020). *Race in America*. W.W. Norton, Incorporated. https://books.google.com/books?id=03ThxwEACAAJ

Devos, T., & Banaji, M. R. (2005). American = White? *Journal of Personality and Social Psychology, 88*(3), 447–466. https://doi.org/10.1037/0022-3514.88.3.447

Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology*, *8*(2), 88–102. https://doi.org/10.1037/1099-9809.8.2.88

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "Minimal" Group Affiliations in Children. *Child Development*, *82*(3), 793–811. https://doi.org/10.1111/j.1467-8624.2011.01577.x

Dunning, D. (2007). Self-Image Motives and Consumer Behavior: How Sacrosanct Self-Beliefs Sway Preferences in the Marketplace. *Journal of Consumer Psychology*, *17*(4), 237–249. https://doi.org/10.1016/S1057-7408(07)70033-5

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes. *Psychological Science*, *17*(5), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Effron, D. A., Miller, D. T., & Monin, B. (2012). Inventing racist roads not taken: The licensing effect of immoral counterfactual behaviors. *Journal of Personality and Social Psychology*, *103*(6), 916–932. https://doi.org/10.1037/a0030008

Embrick, D. G. (2011). The Diversity Ideology in the Business World: A New Oppression for a New Age. *Critical Sociology*, *37*(5), 541–556. https://doi.org/10.1177/0896920510380076

Emde, R. N., Biringen, Z., Clyman, R. B., & Oppenheim, D. (1991). The moral self of infancy: Affective core and procedural knowledge. *The Development of Self: The First Three Years*, *11*(3), 251–270. https://doi.org/10.1016/0273-2297(91)90013-E

Freeman, L., & Stewart, H. (2021). Toward a Harm-Based Account of Microaggressions. *Perspectives on Psychological Science*, *16*(5), 1008–1023. https://doi.org/10.1177/17456916211017099

Friedlaender, C. (2020). A Defense of Intentional Microaggressions and Microaggressive Harassment. *Microaggressions and Philosophy*, 184–204. https://doi.org/10.4324/9780429022470-9

Frimer, J. A., & Walker, L. J. (2009). Reconciling the self and morality: An empirical model of moral centrality development. *Developmental Psychology*, *45*(6), 1669–1681. https://doi.org/10.1037/a0017418

Frisch, L. K., Kneer, M., Krueger, J. I., & Ullrich, J. (2021). The effect of outcome severity on moral judgement and interpersonal goals of perpetrators, victims, and bystanders. *European Journal of Social Psychology*, *51*(7), 1158–1171. https://doi.org/10.1002/ejsp.2805

Gaertner, S. L., & Dovidio, J. F. (2005). Understanding and Addressing Contemporary Racism: From Aversive Racism to the Common Ingroup Identity Model. *Journal of Social Issues*, *61*(3), 615–639. https://doi.org/10.1111/j.1540-4560.2005.00424.x

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel. *Psychological Science*, *20*(3), 393–398. https://doi.org/10.1111/j.1467-9280.2009.02306.x

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless+harmless=blameless: When seemingly irrelevant factors influence judgment of (un)ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93–101. https://doi.org/10.1016/j.obhdp.2009.11.001

Gray, K., & Wegner, D. M. (2011). To escape blame, don't be a hero—Be a victim. *Journal of Experimental Social Psychology*, *47*(2), 516–519. https://doi.org/10.1016/j.jesp.2010.12.012

Hammack, P. L. (2011). Narrative and the politics of meaning. In *Narrative Inquiry* (Vol. 21, Issue 2, pp. 311–318). John Benjamins. https://doi.org/10.1075/ni.21.2.09ham

Hardy, S. A., & Carlo, G. (2011). Moral Identity: What Is It, How Does It Develop, and Is It Linked to Moral Action? *Child Development Perspectives*, *5*(3), 212–218. https://doi.org/10.1111/j.1750-8606.2011.00189.x

Heiphetz, L. (2019). Moral essentialism and generosity among children and adults. *Journal of Experimental Psychology: General*, *148*(12), 2077–2090. https://doi.org/10.1037/xge0000587

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The Role of Moral Beliefs, Memories, and Preferences in Representations of Identity. *Cognitive Science*, *41*(3), 744–767. https://doi.org/10.1111/cogs.12354

Heiphetz, L., & Young, L. (2014). A social cognitive developmental perspective on moral judgment. *Behaviour*, *151*(2–3), 315–335. https://doi.org/10.1163/1568539X-00003131

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral Punishment in Everyday Life. *Personality and Social Psychology Bulletin*, *44*(12), 1697–1711. https://doi.org/10.1177/0146167218775075

Isbell, L. M., & Tyler, J. M. (2003). Teaching Students About In-Group Favoritism and the Minimal Groups Paradigm. *Teaching of Psychology*, *30*(2), 127–130. https://doi.org/10.1207/S15328023TOP3002_10

Janneck, M., Bayerl, P. S., & Dietel, J.-E. (2013). The Minimal Group Paradigm in Virtual Teams. In A. Holzinger, M. Ziefle, M. Hitz, & M. Debevc (Eds.), *Human Factors in Computing and Informatics* (pp. 457–476). Springer Berlin Heidelberg.

Johnston, M., & Krettenauer, T. (2011). Moral self and moral emotion expectancies as predictors of anti- and prosocial behaviour in adolescence: A case for mediation? *European Journal of Developmental Psychology*, *8*(2), 228–243. https://doi.org/10.1080/17405621003619945

Jordan, J., & Kouchaki, M. (n.d.). Virtuous victims. *Science Advances*, *7*(42), eabg5902. https://doi.org/10.1126/sciadv.abg5902

Jordan, J., Leliveld, M. C., & Tenbrunsel, A. E. (2015). The Moral Self-Image Scale: Measuring and Understanding the Malleability of the Moral Self. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01878

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior. *Personality and Social Psychology Bulletin*, *37*(5), 701–713. https://doi.org/10.1177/0146167211400208

Joseph, P. (2017). *The New Human Rights Movement: Reinventing the Economy to End Oppression*. BenBella Books. https://books.google.com/books?id=TrhCDgAAQBAJ

Kagan, S. (1998). *Normative Ethics* (Vol. 109, Issue 434, pp. 373–377). Westview Press.

Keller, M., Lourenço, O., Malti, T., & Saalbach, H. (2003). The multifaceted phenomenon of 'happy victimizers': A cross-cultural comparison of moral emotions. *British Journal of Developmental Psychology*, *21*(1), 1–18. https://doi.org/10.1348/026151003321164582

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197–215. https://doi.org/10.1016/j.cognition.2011.01.006

Kneer, M., & Machery, E. (2019). No luck for moral luck. *Cognition*, *182*, 331–348. https://doi.org/10.1016/j.cognition.2018.09.003

Kneer, M., & Skoczeń, I. (2023). Outcome effects, moral luck and the hindsight bias. *Cognition*, *232*, 105258. https://doi.org/10.1016/j.cognition.2022.105258

Knowles, E. D., Lowery, B. S., Chow, R. M., & Unzueta, M. M. (2014). Deny, Distance, or Dismantle? How White Americans Manage a Privileged Identity. *Perspectives on Psychological Science*, *9*(6), 594–609. https://doi.org/10.1177/1745691614554658

Kochanska, G. (2002). Committed compliance, moral self, and internalization: A mediational model. *Developmental Psychology*, *38*(3), 339–351. https://doi.org/10.1037/0012-1649.38.3.339

Lowery, B. S., Knowles, E. D., & Unzueta, M. M. (2007). Framing Inequity Safely: Whites' Motivated Perceptions of Racial Privilege. *Personality and Social Psychology Bulletin*, *33*(9), 1237–1250. https://doi.org/10.1177/0146167207303016

Lui, P. P., Berkley, S. R., Pham, S., & Sanders, L. (2020). Is microaggression an oxymoron? A mixed methods study on attitudes toward racial microaggressions among United States

university students. *PLOS ONE*, *15*(12), e0243058. https://doi.org/10.1371/journal.pone.0243058

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133–143. https://doi.org/10.1016/j.cognition.2015.11.008

Masek, L. (2010). Intentions, Motives, and the Doctrine of double effect. *The Philosophical Quarterly*, *60*(240), 567–585. https://doi.org/10.1111/j.1467-9213.2009.633.x

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

McCabe, J. (2009). Racial and Gender Microaggressions on a Predominantly-White Campus: Experiences of Black, Latina/o and White Undergraduates. *Race, Gender & Class*, *16*(1/2), 133–151. JSTOR.

McMahan, J. (1994). Revising the Doctrine of Double Effect. *Journal of Applied Philosophy*, *11*(2), 201–212. https://doi.org/10.1111/j.1468-5930.1994.tb00109.x

Medwed, D. S. (2020). Black deaths matter: The race-of-victim effect and capital punishment. *Brooklyn Law Review*, *86*(3), 957.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Miller, J. G. (1994). Cultural Diversity in the Morality of Caring: Individually Oriented Versus Duty-Based Interpersonal Moral Codes. *Cross-Cultural Research*, *28*(1), 3–39. https://doi.org/10.1177/106939719402800101

Molenberghs, P., Gapp, J., Wang, B., Louis, W. R., & Decety, J. (2016). Increased Moral Sensitivity for Outgroup Perpetrators Harming Ingroup Members. *Cerebral Cortex*, *26*(1), 225–233. https://doi.org/10.1093/cercor/bhu195

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. *Personality, Identity, and Character: Explorations in Moral Psychology*, 341–354.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, *108*(7), 2688–2692. https://doi.org/10.1073/pnas.1011734108

Mulder, L. B., & Aquino, K. (2013). The role of moral identity in the aftermath of dishonesty. *Organizational Behavior and Human Decision Processes*, *121*(2), 219–230. https://doi.org/10.1016/j.obhdp.2013.03.005

Mummendey, A., Otten, S., Berger, U., & Kessler, T. (2000). Positive-Negative Asymmetry in Social Discrimination: Valence of Evaluation and Salience of Categorization. *Personality and Social Psychology Bulletin*, *26*(10), 1258–1270. https://doi.org/10.1177/0146167200262007

Nadal, K. L., Khogali, M., Nguyễn, P. C., & Erazo, T. (2020). Microaggressions and Implicit Biases: Rooted in Structural Racism and Systemic Oppression. In R. Geffner, J. W. White, L. K. Hamberger, A. Rosenbaum, V. Vaughan-Eden, & V. I. Vieth (Eds.), *Handbook of Interpersonal Violence and Abuse Across the Lifespan: A project of the National Partnership to End Interpersonal Violence Across the Lifespan (NPEIV)* (pp. 1–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-62122-7_101-1

Navarrete, C. D., Kurzban, R., Fessler, D. M. T., & Kirkpatrick, L. A. (2004). Anxiety and Intergroup Bias: Terror Management or Coalitional Psychology? *Group Processes & Intergroup Relations*, *7*(4), 370–397. https://doi.org/10.1177/1368430204046144

Norton, M. I., & Sommers, S. R. (2011). Whites See Racism as a Zero-Sum Game That They Are Now Losing. *Perspectives on Psychological Science*, *6*(3), 215–218. https://doi.org/10.1177/1745691611406922

Okonofua, J. A., & Eberhardt, J. L. (2015). Two Strikes: Race and the Disciplining of Young Students. *Psychological Science*, *26*(5), 617–624. https://doi.org/10.1177/0956797615570365

Payir, A., & Heiphetz, L. (2022a). Children's and Adults' Attribution of Moral Judgments to Human and Supernatural Agents. *Journal of Cognition and Development*, *23*(4), 524–544. https://doi.org/10.1080/15248372.2022.2061975

Payir, A., & Heiphetz, L. (2022b). Children's and Adults' Attribution of Moral Judgments to Human and Supernatural Agents. *Journal of Cognition and Development*, *23*(4), 524–544. https://doi.org/10.1080/15248372.2022.2061975

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Prentice, M., Jayawickreme, E., Hawkins, A., Hartley, A., Furr, R. M., & Fleeson, W. (2019). Morality as a Basic Psychological Need. *Social Psychological and Personality Science*, *10*(4), 449–460. https://doi.org/10.1177/1948550618772011

Quinn, W. S. (1989). Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy & Public Affairs*, *18*(4), 334–351. JSTOR.

Recchia, H. E., Wainryb, C., Bourne, S., & Pasupathi, M. (2015). Children's and Adolescents' Accounts of Helping and Hurting Others: Lessons About the Development of Moral Agency. *Child Development*, *86*(3), 864–876. https://doi.org/10.1111/cdev.12349

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning Saints and Saintly Sinners: The Paradox of Moral Self-Regulation. *Psychological Science*, *20*(4), 523–528. https://doi.org/10.1111/j.1467-9280.2009.02326.x

Schonert-Reichl, K. A., Smith, V., Zaidman-Zait, A., & Hertzman, C. (2012). Promoting Children's Prosocial Behaviors in School: Impact of the "Roots of Empathy" Program on the Social and Emotional Competence of School-Aged Children. *School Mental Health*, *4*(1), 1–21. https://doi.org/10.1007/s12310-011-9064-7

Schug, M. G., Shusterman, A., Barth, H., & Patalano, A. L. (2013). Minimal-group membership influences children's responses to novel experience with group members. *Developmental Science*, *16*(1), 47–55. https://doi.org/10.1111/j.1467-7687.2012.01193.x

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauerschatka, H., & Simons, A. (1991). Ease of Retrieval as Information—Another Look at the Availability Heuristic. *Journal of Personality and Social Psychology, v.61, 195-202 (1991), 61*. https://doi.org/10.1037//0022-3514.61.2.195

Sengsavang, S., Willemsen, K., & Krettenauer, T. (2015). Why be moral? Children's explicit motives for prosocial-moral action. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00552

Sidanius, J., & Pratto, F. (1999). *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge University Press; Cambridge Core. https://doi.org/10.1017/CBO9781139175043

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In *Advances in Experimental Social Psychology* (Vol. 34, pp. 379–440). Academic Press. https://doi.org/10.1016/S0065-2601(02)80009-0

Sticker, R. M., Christner, N., Pletti, C., & Paulus, M. (2021). The moral self-concept in preschool children: Its dimensions and relation to prosocial behaviors. *Cognitive Development*, *58*, 101033. https://doi.org/10.1016/j.cogdev.2021.101033

Sue, D. W., Alsaidi, S., Awad, M. N., Glaeser, E., Calle, C. Z., & Mendez, N. (2019). Disarming racial microaggressions: Microintervention strategies for targets, White allies, and bystanders. *American Psychologist*, *74*(1), 128–142. https://doi.org/10.1037/amp0000296

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, *62*(4), 271–286. https://doi.org/10.1037/0003-066X.62.4.271

Sue, D. W., & Spanierman, L. B. (2020). *Microaggressions in everyday life: Race, gender, and sexual orientation.* John Wiley & Sons, Ltd.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Testa, F. (2023). The boundaries of moral responsibility and racially oriented microaggressions. *Journal of Theoretical and Philosophical Psychology*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/teo0000264

Todd, A. R., & Galinsky, A. D. (2014). Perspective-Taking as a Strategy for Improving Intergroup Relations: Evidence, Mechanisms, and Qualifications. *Social and Personality Psychology Compass*, *8*(7), 374–387. https://doi.org/10.1111/spc3.12116

Tybout, A. M., Sternthal, B., Malaviya, P., Bakamitsos, G. A., & Park, S.-B. (2005). Information Accessibility as a Moderator of Judgments: The Role of Content versus Retrieval Ease. *Journal of Consumer Research*, *32*(1), 76–85. https://doi.org/10.1086/426617

Valdesolo, P., & DeSteno, D. (2007). Moral Hypocrisy: Social Groups and the Flexibility of Virtue. *Psychological Science*, *18*(8), 689–690. https://doi.org/10.1111/j.1467-9280.2007.01961.x

Wainryb, C., Brehl, B. A., Matwin, S., Sokol, B. W., & Hammond, S. (2005). Being Hurt and Hurting Others: Children's Narrative Accounts and Moral Judgments of Their Own Interpersonal Conflicts. *Monographs of the Society for Research in Child Development*, *70*(3), i–122. JSTOR.

Wohl, M. J. A., Branscombe, N. R., & Klar, Y. (2006). Collective guilt: Emotional reactions when one's group has done wrong or been wronged. *European Review of Social Psychology*, *17*(1), 1–37. https://doi.org/10.1080/10463280600574815

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214. https://doi.org/10.1016/j.cognition.2011.04.005

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, *145*(11), 1448–1459. https://doi.org/10.1037/xge0000190