Causes and Consequences of Coalitional Cognition

Mina Cikara
Harvard University

Invited, *Advances in Experimental Social Psychology*

I am very grateful to Joel Eduardo Martinez and David Pietraszewski for their insightful comments on drafts of this paper. This work was generously supported by the National Science Foundation (BCS-1551559 and a CAREER award, BCS-1653188 awarded to MC).

Mina Cikara
Department of Psychology
Harvard University
33 Kirkland St.
William James Hall 1420
Cambridge MA 02138
mcikara@fas.harvard.edu

What is a group? How do we know to which groups we belong? How do we assign others to groups? A great deal of theorizing across the social sciences has conceptualized ‘groups’ as synonymous with ‘categories,’ however there are a number of limitations to this approach: particularly for making predictions about novel intergroup contexts or about how intergroup dynamics will change over time. Here I join a growing chorus of researchers striving to systematize the conditions under which a generalized coalitional psychology gets activated—the recognition of another’s capacity for and likelihood of coordination not only with oneself but with others. First I review some recent developments in the cognitive processes that give rise to the inference of coalitions and group-biased preferences (even in the absence of category labels). Then I review downstream consequences of inferences about capacity and likelihood of coordination for valuation, emotions, attribution, and inter-coalitional harm. Finally I review examples of how we can use these psychological levers to attenuate intergroup hostility.

Causes and Consequences of Coalitional Cognition

"[It] is safe to assume that there cannot be separate psychologies of prejudice in relation to this or that group, but that they are specific cases of the general picture of prejudice." - Sherif, 1948 (p. 64)

1. INTRODUCTION

Groups are a fundamental organizing principle of our psychology and behavior. Our tendency to carve up the world into 'us' and 'them' is the source of our greatest triumphs but also our greatest tragedies. Identifying and coordinating with fellow group members allows people to satisfy their own material and psychological needs (Allport, 1954) and to develop norms and practices that bolster our most cherished social institutions (e.g., Keltner, 2009; Tomasello, 2009). However, group-living is a double-edged sword (Bornstein, 2003; Halevy, Chou, Cohen, & Bornstein, 2010; Yzerbyt & Demoulin, 2010). Even under the most mundane conditions, people prefer in-group members to out-group members, (Brewer, 1979; LeVine & Campbell, 1972; Mullen, Brown, & Smith, 1992; Perdue, Dovidio, Gurtman, & Tyler, 1990), treat in-group members more favorably (Hewstone, Rubin, & Willis, 2002), allocate more resources to in-group members (Tajfel, Billig, Bundy, & Flament, 1971), and exert more effort on behalf of the in-group's goals (Ellemers, De Gilder, & Haslam, 2004). When intergroup relations become more acrimonious, violence and conflict abound (Cohen & Insko, 2008). According to one statistic, more than 170 million civilians were killed in the 20th century by acts of genocide, war, and other forms of group conflict (Woolf & Hulsizer, 2004). Given how consequential groups are and how much time and resources we have dedicated to studying them, it is surprising that a deep understanding of "groups" as a general concept still eludes us (Pietraszewski, in press).

What is a group? How do we know to which groups we belong? How do we assign others to groups? Driven in part by the prominence of social identity theory (SIT; Tajfel & Turner, 1979) and the minimal groups paradigm, the contemporary intergroup literature has emphasized the role of category-membership (e.g., Black vs. White people; Rattlers vs. Eagles) over coalitional structure (i.e., friends vs. foes). This approach is limited, however, because social categories aren't fixed entities (Zárate, Reyna, & Alvarez, 2019). For one, the associations with specific categories change over time (e.g., when Italian and Irish immigrants became 'White' in American in the early 20th century) or categories may fracture forming new categories (Moya & Scezla, 2015). Allegiances *between* categories can change (e.g., compare Americans' relationship to Germans now versus 80 years ago) which naturally changes the nature of attitudes and behavior that unfold in those intergroup contexts. Third, not all categories carry with them the psychological potency of purposive groups (e.g., we have never been concerned about an uprising of Brunettes—but it is 2020 as I write this so we may yet be surprised). Thus studying categories is unlikely to get us very far in the pursuit of understanding the general concept of 'groups.'

Before SIT there were relationship-based theories—specifically theories centered on perceptions of intergroup threat. Realistic Group Conflict Theory proposed that competition for access to limited resources is a central driver of conflict between groups (Levine & Campbell, 1972; Sherif, 1966). Since its introduction, RGCT has been extended to predict that mere *perception* of threat (even in the absence of actual threat) is sufficient to ignite and sustain

conflict (Esses, Jackson, & Armstrong, 1998; Stephan & Stephan, 2017). In complement, Symbolic Threat Theory (Kinder & Sears, 1981) posited that intergroup conflict can also result from conflicting values and beliefs between groups. Broadly, the distinction is that groups who consume resources or threaten one's general welfare pose realistic threats whereas groups whose values and ideologies are at odds with our group's pose symbolic threats. More recent theorizing suggests that threat does not even need to be linked to social identity or groups per se to have significant consequences for intergroup dynamics. The threat may arise from some feature of the environment (e.g., disease, resource scarcity), which then impacts perceptions of out-groups and their members as threats (Schaller & Neuberg, 2012; Krosch, Tyler, & Amodio, 2017). I'll refer to these as coalitional accounts because they speak to the roots of intergroup discrimination and conflict invariant to the groups in question.

There is long standing debate regarding whether prejudice and conflict is better accounted for by a category-based account or a coalitional account (Scheepers, Spears, Doosje, & Manstead, 2006; Yamagishi et al., 1999). However I'd argue which approach makes the most sense depends largely on one's aims. If one's goal is to maximize variance explained in an outcome within a particular intergroup context in a particular space and time, then one may be best served by accounting for the particulars of the groups in question (e.g., specific histories between groups, relative group status/power in that moment, stereotypes associated with each group, respectively). However, if one's goal is to make more generalized predictions about novel intergroup contexts or about how intergroup dynamics—even among specific, existing groups—may change going forward, one may be better served by identifying the contextual features and psychological interdependencies that imbue collections of individuals with the status of purposive groups or coalitions (Balliet et al., 2017; De Dreu, Gross, Fariña Ma, 2020; Pietraszewski, Tooby, Cosmides, 2014; Pietraszewski, 2016; 2020; Yamagishi & Kiyonari, 2000). It is noteworthy that in even in illustrations of purportedly "pure" category-based discrimination there are coalitional features at play. For example, the resource allocation task used in the classic Tajfel et al. (1971) minimal groups paradigm had built into it interdependence across participants, because each person's payout was determined by fellow in-group and out-group members' behavior. As such, many have countered that even some minimal group findings can be re-cast as byproducts of a coalitional calculus (Gaertner & Schopler, 1998; Rabie Schot, & Visser, 1989; Yamagishi & Kiyonari, 2000).

Here I join this growing chorus of researchers striving to systematize the conditions under which coalitional psychology gets activated: the recognition of another's capacity for and likelihood of coordination not only with oneself but with others. More specifically, by detection of coordination I mean **believing that another is able and willing (or not) to engage in behavior, taking into account your and others' welfare**[1]. First I review some recent developments in the cognitive processes that give rise to the inference of coalitions and group-biased preferences (even in the absence of category labels). Then I review downstream consequences of inferences about capacity and likelihood of coordination for valuation,

[1] Note that this is more specific than merely having common goals. For example, we may have similar preferences that give rise to a common goal: we both like cookies so we'll work together to find some. But what if we're only able to find one cookie? Now the resource becomes a source of competition which may easily dismantle our cooperative stance. Coordination includes within it expectations of considerations of one another's welfare and corresponding reciprocity (Yamagishi & Kiyonari, 2000).

emotions, attribution, and inter-coalitional harm. Finally I review examples of how we might use these psychological levers to attenuate intergroup hostility.

2. CAUSES

2.1 Moving away from categories toward coalitions

Before any consequences of coalitional cognition can be rendered, people have to identify others as in-group or out-group members. This process of social categorization is unique compared to non-social categorization because social categorization and *identification* are intimately intertwined—we do not just sort people into categories the way we do fruits and vegetables, we sort them into in-groups and out-groups, which are egocentrically defined: those to which *I* do or do not belong.

Thus one process by which people may determine whether someone is an in-group member is via judgments of similarity to one's self on a feature that is relevant to the current context. Among the most widely studied features are skin tone, language, nation of origin, display of symbols signaling religious or sports team affiliation, and for good reason. These features can become associated with *expectations* of capacity for and likelihood of coordination (or conflict) not only with oneself but with others (Pietraszewski, 2013).

There are, however, many cases where shared category-membership fails to predict coordination. One such example is the Queen Bee effect (Staines, Tavris, & Jayaratne, 1974). This refers to a phenomenon in which successful women not only fail to promote other people who share a female identity, they are particularly punitive toward them. One explanation for this pattern is that the successful women had to fight negative stereotypes about femininity to ascend to their current ranks, and so they come to denigrate that aspect of their identity—for example, they characterize themselves in more masculine terms (for review see Derks, Van Laar, Ellemers, 2016). They subsequently become active gatekeepers against those who don't denounce their feminine identity—for example being even less likely to support the hiring or promotion of female-identifying colleagues (e.g., Derks, Ellemers et al., 2011; Derks, Van Laar et al., 2011). This example highlights one of the limitations in terms of thinking about social categories as meaningful groups; because people within a given category can, by virtue of surviving in these structures, become the greatest weapon against their fellow category-group members. Nor is this phenomenon unique to gender (Huddy, 2001): consider as other examples, Hispanic/Latinx Immigrations and Customs Enforcement officers, the adage "not all skinfolk is kinfolk," or widespread confusion as to how so many Latinx voters supported Trump in the 2020 election. Research treating demographic categories as purposive groups will often run into these explanatory limitations *and* may ironically end up reinforcing stereotypes and the belief that categories are social monoliths (Brick, Hood, Ekroll, & de-Wit, in press; Martinez & Paluck, 2020). If similarity is a poor cue on which to base group inferences, on what else might we rely?

Classic social psychological theories of group perception and entitativity[2] remind us that in addition to similarity there are several other dimensions by which groups may be defined,

[2] While the work on entitativity is central to this inquiry, I argue that it lacks predictive power. Dimensions like similarity and proximity give rise to increased attributions of entitativity but where do judgments of, for example, similarity and common fate come from? Similarity on what dimensions? Is common fate an antecedent to or consequence of grouping? All of the different types of groups that have been characterized by work on entitativity—e.g., intimacy, task, or

including common fate within groups (Campbell, 1958). Common fate refers to conditions under which individual group members' outcomes are interdependent (e.g., group members share exposure to threat or benefit). Not only does common fate increase perceptions of group cohesion within groups (Hamilton & Sherman, 1996), it promotes greater intergroup bias and discrimination between groups (Gaertner & Schopler, 1998); therefore common fate appears to be a relatively stronger cue to group inference than similarity. Indeed, when similarity and common fate are pit against one another, common fate is a better predictor of behavior. For example, when group member similarity, proximity, and common fate are independently manipulated, common fate is the only significant predictor of competitive, group-based aggression in the prisoner's dilemma game (Insko, Wildschut, & Cohen, 2013). To be clear, forces such as common fate can be and often are martialed to imbue mere categories with coalitional potency. For example, when a class of people are systematically subjugated and oppressed by a dominant class (e.g., the oppression and brutalization of Black Americans by White Americans) their category takes on coalitional status via their common fate at the hand of the oppressing class, but not because there is something intrinsic to the category, e.g., 'Black people,' that makes it more monolithic relative to other demographic categories (see related work on shared experience of discrimination binding multiple minoritized groups; Cortland et al., 2017).

Thus feeling that one's outcomes are tied to another's at the hand of some outside force (social or otherwise, e.g., a hurricane) can act as a cue to shared group membership but speaking more specifically to the importance of coordination, people also have strong expectations about the nature of the *interactions and the obligations* within and between group members: specifically that people within groups will cooperate or try harder to coordinate with one another while reserving obstruction, competition, and harm for out-group members (Balliet et al., 2017; De Dreu et al. 2020; Kelley & Thibaut, 1978; Pietraszewski, in press; Rhodes & Chalik, 2013; Yamagishi & Mifune, 2016). Perhaps not surprising then is that even very young children prefer characters who help others to achieve their goals over those who obstruct those goals (Hamlin, Wynn, & Bloom, 2007; Kuhlmeier, Wynn, & Bloom, 2003). Note, however, that children only prefer these characters if they helped intentionally and with knowledge of their interaction partner's preferences, suggesting that what children are tracking is the *coordination* between one character's desires and the other character's response in service of those desires (Hamlin et al., 2013). In line with this, children use resource sharing as a strong cue to social allegiances: 4 to 9-year old's generally believe distributors who give more to agent A than agent B are better friends with A than B (Liberman & Shaw, 2017). Rather than being outputs of group membership (as they are framed in the minimal groups' paradigm) these coordination cues may be the primary inputs to group inference.

Even cues that are often classified as "arbitrary characteristics" may signal important information about capacity or willingness to coordinate. For example, language is often cited as an arbitrary characteristic that guides children's social preferences—children prefer people who speak the same language or share their accent (Kinzler, Dupoux, & Spelke, 2007). For young

---

loose association groups (Lickel et al., 2000)—can be animated to triumph or atrocity (e.g., people waiting at the bank, a loose association, may be mobilized by an armed robbery) just like any of them can be neutralized to indifference. We need a unified theory that can generalize across these levels and examples and account for the features of the context that give rise to purposive groups.

children (as well as adults), though, language is anything but arbitrary; communication is an act of coordination. Said another way, not being able to communicate drastically reduces the possibility of successful coordination. In line with this characterization, partner accent trumps other category features, including skin tone, in driving children's social preferences (Kinzler, Schutts, DeJesus, & Spelke, 2009). Furthermore, accent, like gender, cannot be overridden by competing coalitional information the way that race and ethnicity (i.e., skin tone) can (Pietraszewski & Schwartz, 2014).

Thus even in the absence of threat or competition, the inference of coordination difficulty or improbability may be sufficient to mark someone as an out-group member. Perhaps more important, and as noted above, coordination ease (or difficulty) may or may not track with shared category membership. Finally, people, including very young children, are sensitive to how well agents coordinate not just with themselves but with others in the environment, indicating that people are prone to building representations of coordinated coalitions—or social structures—out in the world rather than just egocentric, dyadic similarities or interdependencies. (As an example of how early emerging these inter-agent coordination sensitivities are, even *infants* prefer third-party characters who obstruct hinderers and assist helpers; Hamlin et al., 2011.)

We set out to test what information people use to build these coalitional representations. Specifically we tested to what extent people rely on similarity versus latent group structure, based on observable behavior, to guide their choices of allies in the absence of any category labels. A mere similarity account predicts that people simply substitute judgments of similarity to the self on relevant features (e.g., did this person vote for the same candidate I did in the last election?) to identify allies. An alternative hypothesis is that people's inferences about coalition membership may be further improved by integrating information both about how agents relate to oneself as well as how they relate to one another (e.g., "How well do I get along with Mina? With Carey? How do they get along with each other?"; Heider, 1958).

We have recently proposed a formal account of social group discovery in which we adopt a computational model of latent structure learning to move beyond explicit category labels and dyadic similarity as the sole inputs to social group or coalition representations (Gershman & Cikara, 2020). Specifically, we examine whether people in addition to tracking dyadic similarity also build representations of latent groups in the environment via Bayesian inference. If people represent latent group structure in addition to dyadic similarities, then even when two agents' choices are equally similar to their own, their decisions should be influenced by the presence of a third agent that alters the coalitional structure. Importantly, prior models that rely on dyadic similarity would not predict differential social influence in these cases (because similarity is equated for the first two agents in question).

In a series of behavioral experiments framed as learning about strangers' political issue positions, we tested whether the degree to which participants were willing to align with one of two agents was affected by the presence of a third agent, who formed a cluster that either did or did not include the participant (Lau, Pouncy, Gershman, & Cikara, 2018; see Figure 1, below). On each trial, participants stated their position for or against a political issue, and then predicted the choices of three other individuals on that same issue. After each prediction, they received feedback about that individual's actual choice. Finally, at the end of this learning phase, participants had to choose with which agent—Agent A or B—they wanted to align themselves on a "mystery issue." Critically, Agents A and B agreed with the participant an equal number of times, making them equally similar to the participant. Depending on the block, however, Agent C either clustered with Agent B *and* the participant, or only with Agent B and not the participant.

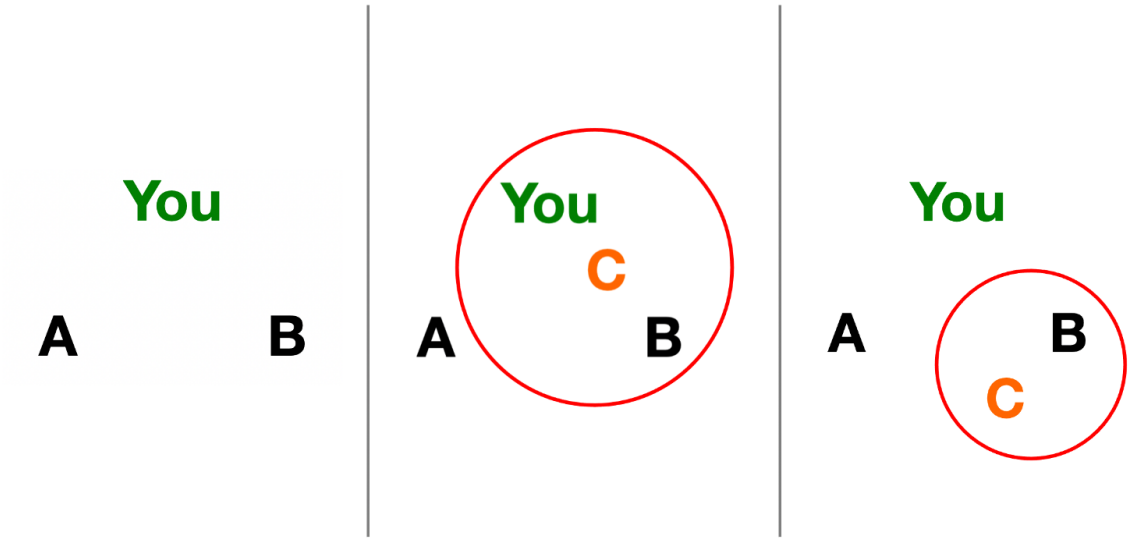

Figure 1. Schematic representation of different coalitional structures as a function of Agent C. In these figures distance is a proxy for similarity. In all panels A and B are equally similar to you, but in the middle panel C's placement creates a group that includes both you and B (which increases your preference for B relative to A), whereas in the right panel C's placement puts B in a group that does not include you. Adapted from Lau et al. (2018).

As predicted by a latent structure learning account, participants favored Agent B over Agent A when C's placement created a cluster that put the participant in the same group as Agent B (despite the fact that Agents A and B were equally similar to the participant). Furthermore, the influence of the latent social groups generalized to other judgments. Participants also judged Agent B as more competent, moral, and likable than Agent A when Agent B clustered with the participant versus not. Perhaps most interesting, latent structures continued to exert an effect on ally-choice behavior even when we provided participants with explicit group labels that contradicted the latent structure (i.e., always put Agent B in the explicit out-group).

In a companion fMRI study, similarity and latent structure learning were associated with distinct neural substrates. Replicating an abundance of earlier work on judgments of similarity to self and categorization of same-category members (e.g., Denny et al., 2012; Jenkins & Mitchell, 2011; Molenberghs & Morrison, 2012; Morrison et al., 2012), trial-by-trial estimates of 'allyship,' based on similarity between participants and each individual agent, recruited vmPFC and pregenual anterior cingulate (pgACC; Figure 2, top). Latent social group structure-based allyship estimates, on the other hand, recruited right anterior insula (rAI; Figure 2, bottom). This specific region within AI overlapped with a region identified by a *non-social* structure learning task (Tomov, Dorfman, & Gershman, 2017), suggesting that it supports domain-general structure representation. Most interesting, however, was that variability in the brain signal from rAI improved prediction of variability in ally-choice behavior, whereas variability from the pgACC did not (Lau, Gershman, & Cikara, 2020). Said another way, the neural signals associated with 'coalitional structure' representations further explained ally-choice behavior whereas 'inter-agent similarity' representations did not.

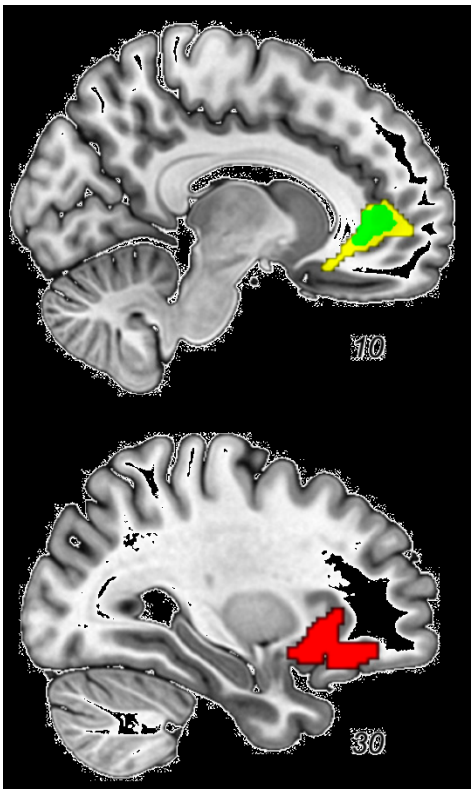

Figure 2. Results from whole-brain contrast (FWE-corrected p<0.05) of parametric modulators. Dyadic similarity model (green), a related *feature* similarity model (yellow), and the latent structure model (red). Top: pgACC (at x = 10); Bottom: rAI (at x = 30). Adapted from Lau et al. (2020).

Therefore, we have rapidly accumulating evidence to indicate that judgments of another's capacity for and likelihood of coordination not only with oneself but with others drives social preferences and choice, both in the absence of category labels and in spite of them.

2.2 Inter-collective functional relations as cues to coalitions

Moving beyond dyadic or triadic interactions, functional relations between *collectives*—whether sets of people are driven by environmental forces to cooperate, compete, or operate independent of one another—are powerful arbiters of coalitional boundaries. For example, cooperation between individuals may (temporarily) change representations of previously marked out-group members to super-ordinate in-group members (Gaertner et al., 2000; Sherif et al., 1961). Thus one hypothesis is that people should be more sensitive to cues to functional relations than same- versus different-category membership. Indeed, there is evidence from both behavioral (Pietraszewski, Cosmides, & Tooby, 2014; Pietraszewski, in press b) and neuroimaging experiments that this is the case.

The first wave of neuroimaging studies of social categorization examined participants' neural responses to same- versus different-category targets (e.g., demarcated by skin tone; for reviews see Amodio & Cikara, 2020; Ito & Bartholow, 2009; Kutbota, Banaji, & Phelps, 2012). However, these early investigations of demographic, category-based groups afforded only limited inferences because the categories were perfectly confounded with differences in the visual appearance of targets, associated stereotypes and prejudices, perceivers' familiarity with

the groups in question, and so on. These confounds make it difficult to infer from the findings which aspects of the results are specific to the categories under investigation versus reflecting generalized group processes.

One obvious follow up was to directly compare the influence of category versus coalitional information. With this objective in mind, Van Bavel et al. (2008) assigned White participants to a minimal mixed-race team consisting of Black and White members and then had them memorize which people were on their own team versus a different team. Participants then viewed Black and White, own-team and other-team faces in the scanner. In contrast to the results documented in studies of neural responses to same-race versus other-race faces—increased amygdala activity to other-race faces—these participants exhibited greater amygdala activity in response to *same-team* faces. There was no main effect of race, nor was this pattern of own-team bias moderated by target race. Coalitional information dominated. However this work was still limited in that it was restricted to one coalition (i.e., the Tigers or Lions teams). In the world, which coalition or group is salient at any given moment is highly context dependent (Ellemers, Spears, & Doosje, 2002; Gaertner et al., 1989; Turner et al., 1994; Van Bavel & Cunningham, 2011). Moreover, the variety of categories with which humans affiliate is vast and each of *these* categories vary on myriad dimensions. So how do we get any traction on generalized group processes?

By some accounts, categorizing people by specific social categories is a byproduct of adaptations that evolved for detecting more general coalitional cues: that is, indicators of functional relations amongst collectives (Pietraszewski et al., 2014; 2020; Sidanius & Pratto, 2001). One thing one would expect to see if this were the case, is a system of representation distinguishing in-group versus out-group members, invariant to the groups in question. Humans would have to have a flexible, common neural code for learning about and representing 'in-group' and 'out-group' targets, invariant to the particular social category or features along which group boundaries are drawn (for review, see Cikara & Van Bavel, 2014). On what brain regions would a common neural code rely? More importantly, what would be the primary structure of the code (e.g., in-group vs. everyone else, threatening out-group vs. everyone else, distinct codes for in-group, neutral out-groups, and threatening out-groups)?

To adjudicate among these competing organizational structures, we conducted an fMRI study that used multi-voxel pattern analysis (MVPA) to test whether participants' neural responses associated with thinking about teammates versus competitors (novel teams, created in the lab) could be used to successfully decode whether they were thinking about political partisans versus opposition (an unrelated, real-world coalition; Cikara, Van Bavel, Ingbretsen, & Lau, 2017). Unlike traditional univariate analysis, MVPA allows investigators to examine different patterns of neural activation within a specific brain region—which may have the same mean-level of activation and thus go undetected by traditional univariate analysis—and use this to distinguish separate psychological representations. In this study, we trained a classifier to encode how people represented the novel teams (i.e., Rattlers vs. Eagles vs. Bears) and then tested how well the neural data decoded membership along a different coalition: political parties (i.e., Democrats vs. Republicans vs. Constitutionals). Any region that results in successful cross-category classification could be said to be representing the higher-order concepts of "us" vs. "them."

Across these coalitions, only two regions were associated with successful cross-categorization: the dorsal ACC/middle cingulate cortex and anterior insula (AI). Interestingly, the dACC and AI are hubs in the 'salience network,' which focuses attention on the most

relevant among internal and external stimuli (both social and non-social) in service of selecting the most sensible behavioral response (e.g., freeze, fight, flight; Menon & Uddin, 2010). This pattern of neural representation associated with the in-group is consistent with the hypothesis that salience, specifically functional significance or evaluation (i.e., will this stimulus help me or not?), is the primary dimension distinguishing representations of us and them (Fiske, Cuddy, & Glick, 2007; Fiske, 2018).

More important, this analysis revealed the structure of this neural code: classification accuracy across coalitions was driven predominantly by the correct classification of in-group targets, consistent with theories indicating in-group identity and expectations of coordination are more central than out-group processing to group perception and cognition (Balliet et al., 2014; Brewer, 1999). It is worth highlighting two points. First, we did not identify a substrate that differentiated between threatening vs. non-threatening out-group representations, despite the fact that people are capable of distinguishing among different out-groups; the primacy of "in-group" observed here may be driven in part by the specifics of this experimental design. Second, the correct categorization of in-group targets was explained in part by the classifier guessing "in-group" more often than any other label across all trials. This over-inclusion pattern makes sense from a statistical learning perspective. People tend to interact more often with people who belong to the same coalitions (McPherson, Smith-Lovin, & Cook, 2001) therefore it us unsurprising that their priors would bias them to assume other people belong to their coalition until evidence indicates otherwise. It is also important to highlight that even though the coalitions featured in these experiments are competitive, they are not associated with threats to individuals' physical safety in any immediate sense. Running the same experiment in social contexts characterized by threats to physical safety (e.g., Gaza) could very well yield an over-*exclusion* bias.

Nevertheless, these findings bolster the notion that there is some basic scaffolding underlying specific inter-coalition interactions on which the details of the relevant categories and groups get overlaid. Understanding the structure of this basic scaffolding helps to reveal which features (e.g., expectations of coordination, absence of competition) organize these representations more generally.

2.3 Addressing some challenges for models of coalitional cognition

Thus one way to think about coalitional cognition is that our main priority is to preserve ourselves and current coalition members (Balliet et al., 2014; De Dreu et al., 2020) and to be vigilant to any indicators of threat to the coalition's security. Said another way, we're not anti-specific out-groups, we're anti-threat (Chang, Krosch, & Cikara, 2018). To state the obvious, the actual or implied presence of threat or competition should lead to low probability estimates of desire to coordinate. The moment feelings of threat are activated, they highlight inter-coalitional boundaries, thereby increasing out-group negativity, out-group homogenization, and in-group solidarity (McDoom, 2012). This principle is the foundation of the functional approach that characterizes Realistic Conflict Theory (Campbell, 1965), Intergroup Threat Theory (Stephan & Stephan, 2017), Social Dominance Theory (Sidanius & Pratto, 2001), and the Stereotype Content Model (Fiske et al. 2002), to name only a subset of the relevant frameworks. In these frameworks, there is nothing intrinsic to specific categories that make them threatening, rather it is (perception of) competition over resources or clashes between value systems that initiate the "othering" process (Brandt & Crawford, 2020; Esses, Jackson, Dovidio, Hodson, 2005).

However, any model of coalitional psychology will have to answer several questions which existing accounts do with varying success. First, coalitional boundaries are often

egocentrically defined ("how well will these other people coordinate with *me* and others in my coalition?"), so how is it that widely-shared category stereotypes emerge across individuals and coalitions? Second, a coalitional account would predict that stereotypes should change as functional relations change among coalitions, so why is it that stereotypes seem so intransigent? Finally, *who* counts as a threat? There are many more people outside of our coalitions than there are in them; at what point does a collective reach a threshold that then marks them as a threat? Below I address each of these questions in turn.

First, how do widely-shared stereotypes of categories emerge if coalitions are egocentrically defined and different coalitions have different enemies? Recent exciting developments in theorizing about stereotype content indicate that there is actually relatively less consensus across people in attributions of warmth to social groups—including how threating groups are—relative to judgments about how agentic or conservative-progressive those same groups are (Koch et al., 2020). In other words, judgments of specific categories' threat, trustworthiness, and honesty are *personal*. Thus while some aspects of stereotypes do appear to be widely-shared (e.g., how powerful members of a category are judged to be), perceptions of threat are relatively more idiosyncratic, as predicted by a coalitional account.

Second why are stereotypes and other category-associated features so impervious to change? On the contrary, the data indicate that race and nationality based stereotypes *have* changed over time—in particular by increasingly omitting the negative stereotype content (Bergsieker, Leslie, Constantine, & Fiske, 2012). Now one possibility is that these stereotype shifts merely reflect shifts in social norms surrounding prejudice expression, as predicted by the group norm theory of attitudes (Sherif & Sherif, 1953). Under this framework people are as prejudiced as they believe they are allowed to be. In fact, in some studies, the correlation between perceptions of acceptability of prejudice toward a category and self-reported feeling thermometer scores for that category exceed .9 (Crandall et al., 2002)! There is mounting evidence of this same flexibility on much shorter timescales even in people's implicit associations (Cone, Mann, & Ferguson, 2017). For example, if participants receive just one extremely negative piece of information about a target ("Bob mutilated a small, defenseless animal") participants' implicit evaluations become significantly more negative relative to their baseline evaluations (this effect was not as marked for extremely positive information; e.g., donating a kidney; Cone & Ferguson, 2015). More generally, intergroup socio-cognitive phenomena like implicit biases (Ofosu, Chambers, Chen, & Hehman, 2019; Payne et al., 2017), and stereotypes (Oliver & Mendelberg, 2000) vary and cluster by spatial geography suggesting that local dynamics strongly inform their valence and content consistent with a coalitional account.

Note that this same flexibility also manifests in behavior and policy. For example, during the 2008 Democratic presidential primary process, Hillary Clinton and Barack Obama supporters gave more money in a dictator game to strangers who supported the same primary candidate (coalitional in-group members) as compared with the rival candidate (out-group members). Two months later after President Obama clinched the nomination, supporters of both candidates coalesced around the party nominee, this bias disappeared (Rand et al., 2009). On a longer time scale, we can look to treatment of Asian immigrants and Asian Americans in the U.S. over the last 150 years. Though contemporary views of Asian people are relatively positive as compared to other minoritized categories (e.g., Waters & Eschbach, 1995), this favorability is a recent development. People often forget that the first major immigration restriction passed in the U.S. was the Chinese Exclusion Act of 1882. The Immigration and Nationality Act of 1952 marked a

loosening of restriction, but nevertheless established annual immigration quotas of no more than 100 people for all nations including those in Asia. By contrast to 70 years ago, Asian immigrants and Americans now comprise one of the fastest-growing groups in the U. S. (Xie & Goyette, 2004; Humes, Jones, & Ramirez, 2011). What changed?

2.4 New insights from a coalitional perspective: Social group reference dependence

That stereotypes and policies can change does not tell us *why* they change. When does a category shift from threatening to neutral? Or from neutral to ally? This is where the coalitional perspective can yield new insights, both accounting for past patterns and making novel predictions. Whether a collective is deemed inside or outside coalitional bounds depends on perceivers' estimates of that collective's ability and intent to engage in coordination or conflict. That is, people should be sensitive to generalized group features that signal threat or coordination, invariant to the groups in question. However, a critical point that is missing from existing coalitional accounts is that estimates of ability and intent will be *reference dependent*.

Reference dependence refers to the phenomenon by which decision-makers' preferences for each option in a choice set shift in predictable ways as a function of the available alternatives (Huber, Payne, & Puto, 1982; Simonson, 1989). These shifts are well documented in consumer behavior contexts: for example, the decoy effect, in which introducing a third inferior product changes consumers' preferences for two original products (Pettibone & Wedell, 2000). More recently we have documented similar preference shifts in social contexts (e.g., Chang, Gershman, & Cikara, 2019). For example, in the context of a hiring decision, we have demonstrated that participants had systematically different preferences for the exact same candidate as a function of the other candidates in the choice set and the salience of the candidate attributes under consideration (Chang & Cikara, 2018). The same logic applies to collective-level inferences—how favorably I feel toward collective A depends not only on my estimates of how likely A is coordinate with my collective, but also how likely B, C, and D are to coordinate with us (and one another) as well.

Here I introduce the concept of ***social group reference dependence***: how we feel about a particular collective or category depends in large part on whether and which other categories are around. Said another way, the heterogeneity of our social ecologies and relative rankings of other collectives within it will affect our attitudes toward each constituent collective. Importing the concept of reference dependence as a driver of coalitional structure can help us get traction on how category-associated prejudice and discrimination will change over time. Notably reference dependence (in the abstract) is central to multiple theories of social identification including self-categorization theory (Turner et al,. 1994) and optimal distinctiveness theory (Brewer, 1991). In both frameworks, which identity becomes salient or most valued is determined by one's context (i.e., who else is around). However, neither provides a means of making quantitative predictions of which identity or attribute will be made most salient in a given context or precisely how identity salience shifts in response to changes in the environment.

There are multiple historical and contemporary examples of the influence of social group reference dependence. For example, in the Netherlands during the Holocaust, Protestants were more likely to offer aid to Jews than Catholics in Catholic regions; conversely, Catholics facilitated Jews' evasion in Protestant areas. Stated more generally, minority status rather than religious doctrine explained who offered assistance to the collective fleeing genocide (Braun, 2014). In another historical example, evidence from archival data indicates that Irish and Italian immigrants became "American" owing in part to inflows of Black people during the Great

Migration (1915-1930). Specifically, increased settlements of southern-born Black people in northern cities (operationalized at the level of metropolitan statistical areas) were associated with local increases in naturalization rates of Irish and Italian immigrants as well as increased intermarriage between immigrants and native-born White people (Fouka, Mazumder, & Tabellini, 2020). Again, more generally: prejudice and discrimination against formerly threatening categories updated with the introduction and expansion of a new category. Similar patterns persist today when we examine the effects of Hispanic population growth in the U.S. For example, priming White Americans with information regarding the growth of Hispanic population and their resulting political power (as compared to a control prime) made them more likely to judge Asian Americans as allies (e.g., predicting that Asian Americans would vote the same way White Americans have; Craig & Lee, under review). As I noted above, collectives should be sensitive to more generalized group features that signal threat, invariant to the groups in question. One such threat feature that is reflected in all these examples and which has garnered a great deal of attention, particularly with increases in shifting demographics all over the globe (UN World Migration Report, 2020), is group size.

That group size serves as a generalized cue to threat is an older idea in psychology and sociology. Group threat theory predicts that as minority collectives get proportionally larger within a region, the majority will perceive them as more threatening—both in terms of competition for resources and jobs (Blalock, 1957) and in terms of their potential for collective action and political power (Pettigrew, 1957)—and will in turn harbor more negative attitudes toward them. Empirical support for this prediction bears out across multiple inter-category contexts: e.g., White people's attitudes toward Black people in the U.S. (Bobo, 1983; Fosset & Kiecolt, 1989; Glaser, 1994; Quillian, 1996), and citizens' attitudes toward immigrants in the U.S. (Hood & Morris, 1997) and the EU (Schlueter & Scheepers, 2010). More recent work on shifting demographics widely replicates these findings: perceptions that minoritized categories are growing in number, and especially that the former majority will soon represent a minority, are strongly associated with prejudice (Craig, Rucker, & Richeson, 2018). However, this relationship between group size and prejudice is not so straightforward. First, evidence varies depends on the level of analysis: country versus state versus municipality (Hjerm, 2009). Because immigrants, refugees, and resident minority groups are not distributed evenly geographically, different communities should exhibit distinct hierarchies of prejudice across different minority categories. Second, people are quite inaccurate about demographic reality; for example, people reliably overestimate the size of immigrant populations (Transatlantic Trends Survey, 2010 wave; Wong et al., 2012).

The social group reference dependence hypothesis highlights another challenge. Size judgments—of individual objects, of collectives—are reference dependent (Stevens, 2017). That is, one's estimate of the size of a target is determined relative to the other accessible targets (e.g., in a choice set or sampled from memory; see the Ebbinghaus illusion as a striking visual example; Figure 3). Thus new categories may have to surpass a particular threshold in size to register as such. What is that threshold? The minoritized group's size relative to the majority? How fast the minortized group grows?

Combining the U.S. Census of Population data with FBI crime records, we constructed a novel county-group-decade dataset to test a corollary of the social group reference dependence hypothesis: rather than being sensitive to the absolute size of any one minority group, majority groups will be sensitive to minority categories' *relative rank* in size, being most discriminating against which ever category represents the largest local minority, followed by the second-ranking

category and so on (Cikara, Fouka, & Tabellini, 2020). Specifically we tested the hypothesis that hate crimes against a specific racial/ethnic category will increase as that category's size-based rank amongst minorities in a county increases in a given decade. We tested this hypothesis focusing on the U.S. and exploiting variation in group size rank across counties for four minority categories – Black, Hispanic/Latinx, Asian, and Arab populations – between 1990 and 2010. Crucially, we always controlled for the relative size of each group, thereby isolating the rank effect from a more general "size" effect. Our analysis also accounted for any time-invariant factor – observable or unobservable – that is specific to each county (e.g., historical attitudes toward minority groups or local "culture") and to each minoritized category (e.g., group-specific levels of prejudice in the U.S. as a whole). We also accounted for decade-specific shocks that might change White people's behavior toward minoritized categories in general (e.g., economic shocks that might increase Whites' propensity for "scapegoating" against minoritized categories in general).

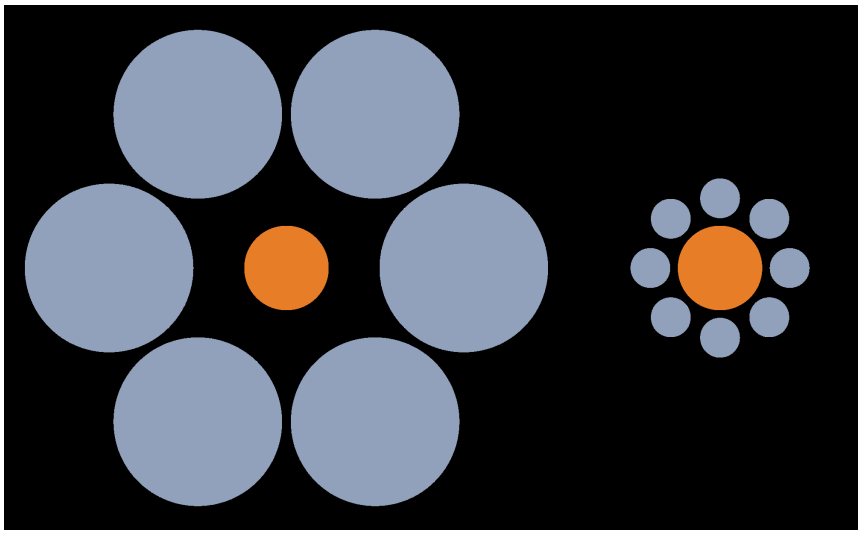

Figure 3. The Ebbinghaus Illusion. The context (i.e., the blue circles) make the orange circle in the center appear larger on the left relative to the right despite the fact that the two orange circles are the same size. Source: https://en.wikipedia.org/wiki/Ebbinghaus_illusion

As predicted, we found that members of the largest minoritized category in a county were significantly more likely to be targeted with hate crimes relative to when their own category ranked second or lower in the minority group size distribution in the same county. This varied by region, so could not be explained by countrywide increase in one category (e.g., Hispanic/Latinx). Furthermore, rank predicted victimization rates independent of whether we measured relative size as a share of total population or of the minoritized population. Third, the rank effect was robust to controlling for the relative category sizes of all other minorities and for the difference in relative size between a minoritized category and the category immediately below it in rank. Finally, rank did not simply capture fast growing minoritized category; our effects remained robust to controlling for the growth rate of each category's size.

Of course, once concern might be that there is some third variable that drives both group size rank and hate crimes. To address this concern, we followed up with an analysis to exploit variation driven by rank *switches* within the same county over time. This strategy compares the change in victimization suffered by two minoritized categories whose relative size, in a given decade, grows by the same amount, but who experience a different "rank change" (e.g. from

second to first in a county versus no change in rank). Even when applying this very stringent empirical test, we find that a category's members experience more victimization in counties where that category's rank moves from second to first relative to counties in which their rank does not change.

Nor were our baseline results restricted only to hate crimes, which represent a rather extreme and relatively rare inter-coalitional behavior. We replicated our analysis of the effect of size rank, this time at the level of the individual, focusing on White respondents' explicit attitudes toward three minority categories: Black, Asian and Arab people (our data source, the Project Implicit database, does not include ratings of Hispanic/Latinx people). Using feeling thermometer ratings as a measure of prejudice, we once again estimated significant effects of category-size rank. Feelings toward the largest minoritized category in a county were significantly more negative compared to the smallest one, conditional on the former's size.

In summary, even though individuals are bad at estimating absolute numbers or even proportions of minority category populations, communities appear to be sensitive to these categories' relative rank in size. This sensitivity is reflected both in county level prejudice as well as extreme manifestations of intergroup hostility. Interestingly, these "rank transformations" of complex data distributions represent a form of efficient-coding present across many domains of decision-making (Bhui & Gershman, 2018) which may help explain why rank outperforms so many other demographic features. I find these results scientifically exciting because they suggest there is promise in the social group reference dependency hypothesis specifically, and the coalitional approach—i.e., an emphasis on generalized coalitional and threat cues for the purposes of theory-building and prediction—more generally.

This framework also makes novel predictions about how demographic shifts may affect coalitional structures going forward. For example, as noted above, in the U.S. Asian Americans continue to be the country's fastest growing racial group, with immigration being a major driver of this growth (U.S. Census Bureau, 2016). As Hispanic/Latinx and Asian populations continue to grow, counties with the greatest relative increases in these populations may see a change in which features matter for prejudice: for example, shifting away from skin tone to language as a primary coalitional boundary. Relatedly, Asian Americans' tenuous status as model minorities (Xu & Lee, 2013) may begin wane as their populations increase, particularly in places where they begin to outnumber other minorities.

A broader theoretical contribution of this framework and these preliminary findings is that they dispel the notion that category attitudes are fixed driven only by essentialized features of the groups themselves. This matters because White people's beliefs about the malleability of racial bias influence their approaches to and strategies within interracial interactions. For example those who (are led to) believe bias is malleable prefer learning oriented strategies (e.g., learning why interracial interactions are challenging) to performance-oriented strategies (e.g., ending the interaction as quickly as possible; Neel & Shapiro, 2012).

Thus in this section we have reviewed the evidence that inferences of individuals' and collectives' capacity and desire for coordination activate coalitional cognition: that is, categorization and ally-choice. Next I turn to come consequences of this activation.

3. CONSEQUENCES

Once coalitional psychology has been activated there is a whole cascade of consequences for our social preferences, emotions, and attributions. However, the nature of these consequences

will hinge on the particulars of the functional relations between coalitions. The absence of cues to coordination may yield indifference, but the overt presence of cues to threat (e.g., competition over resources and incompatibility between groups' goals) give way to conflict (Brewer, 2000) and emotions like fear, hatred, and disgust (Cuddy, Fiske, & Glick, 2007; Mackie & Smith, 2015). Even when coalitions are not explicitly engaged in competition, categories merely stereotyped as competitive (e.g., Asians, professional women; Fiske et al., 2002) may elicit hostile attributions, emotions, and behaviors. These attributions and emotions are often used to justify overt discrimination against and persecution of minoritized groups and their members: for example, propaganda demonizing the Jews in Europe and the Tutsi in Rwanda, as well as anti-miscegenation laws in Nazi Germany and Apartheid South Africa.

In the sections that follow, I will review how judgments of another's capacity for and likelihood of coordination (where perceived competition yields very low estimates of coordination likelihood), not only with oneself but with others, affect social preferences, emotions, and attributions with an emphasis on how each of these, in turn, contribute to increases in willingness to harm.

3.1 Social preferences and values

When people think of intergroup harm they likely bring to mind inter-coalitional skirmishes, genocide, and war. However the accessibility of these exemplars occludes more quotidian manifestations of harm in which typical people participate, for example, every time they cast a ballot. Whether people value some collectives more than others is central to understanding how people resolve social tradeoffs, particularly tradeoffs that help a few at the expense of the many (e.g., welfare policy, healthcare reform). Although most people seem fundamentally opposed (even physiologically averse) to physically harming other people (Cushman et al., 2012; FeldmanHall et al., 2012), many may be quick to abandon their preferences for fairness and moral prohibitions against harm when faced with tradeoffs between more and less valued social categories (e.g., same- versus other-category members, young versus old people; e.g., Awad et al., 2018; Petrinovich et al., 1993; Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009).

We conducted an fMRI study in which we adapted the classic trolley problem to test to what extent fundamental coalitional features—perceived competitiveness and status—account for people's preferences in moral tradeoff scenarios (Cikara, Farnsworth, Harris, & Fiske, 2010). In this dilemma, participants rated how acceptable it was for a third party, Joe, to sacrifice one person in order to save five. Critically, we varied which target was being sacrificed and which targets were being saved (presented in photographs) on each trial. Also important was that we had multiple categories represent each possible combination of perceived competitiveness and status in order to test the effect of the dimensions themselves rather than specific category exemplars: e.g., college students and service people (military, firefighters) to represent low competition/high status, elderly and disabled people to represent low competition/high status, wealthy or high-prestige professional people to represent high competition/high status, and unhomed and drug-addicted people to represent high competition/low status. This last group may seem slightly counter-intuitive: why are unhomed people competitive? This is because of how we defined competition in the study: any resources that go to this target are resources that will not go to me or people like me.

Under the 'warmth primacy hypothesis' (Wojciszke et al., 1998), people should have found it more acceptable to sacrifice high competition targets and to save low-competition

targets because that dimension tracks expectations of *likelihood* of coordination. By contrast, an 'economic valuation hypothesis' predicts that people will engage a cost–benefit analysis of lifetime output potential—that is, *ability*—for sacrificed and saved targets (Lenton, 2002), and therefore find it more acceptable to sacrifice low-status targets to save high-status targets. Ultimately what we found was that these two dimensions interacted to predict social preferences (as predicted by the Stereotype Content Model; Fiske et al., 2002; Fiske et al., 2007). Specifically it was least acceptable to save high-competition, low-status targets and most acceptable to save a group of low-competition, high-status targets.

These findings are particularly stark when we compare them against similar trolley problem data where targets are unspecified. For example, in one finding, 88% of people say that pushing one person off a bridge to save five is unacceptable (Hauser et al., 2007), indicating that most people's default is aversion to sacrifice. However, we *reversed* this pattern by manipulating the perceived competition and status of the targets involved: 84% of our respondents said it was acceptable for Joe to push a high-competition, low-status person off a bridge to save five low-competition, high-status targets. The fMRI results in tandem indicated that left lateral orbital frontal cortex and dorsolateral prefrontal cortex activation was related to this pattern of moral acceptability. Interestingly, some evidence suggests that left lateral OFC is particularly important for the suppression of distressing information and sensations in decision making (Beer, Knight, & D'Esposito, 2006; Bishop, Duncan, Brett, M., & Lawrence, 2004). Thus one possible interpretation of these findings is that participants were actively overriding their moral aversion to using a "low-valued" person as a means to an end when they had the opportunity to save five "high-valued" people.

This application of reduced moral protections to extra-coalitional members is not relegated to hypothetical scenarios—people are reliably more aggressive when they act on behalf of their coalitions relative to when they act alone (Cohen, Montoya, & Insko, 2006; Meier & Hinsz, 2004): even more so when collectives are in explicit competition with one another (Hamilton & Sherman, 1996; Tajfel, 1982). There are several reasons why this may be (Bandura, 1999): aggression may be reframed as virtuous or necessary for a greater cause (e.g., Fiske & Rai, 2014), or collectives may afford a feeling of anonymity or diffusion of responsibility (e.g., Darley and Latané, 1968). A third lesser explored explanation for this pattern is that acting with a collective allows coalitional priorities to supersede one's own personal moral standards. Said another way, people may be less likely to reference their personal moral standards in competitive intergroup contexts relative to when they are acting in their own interests. Testing this hypothesis, however, is quite challenging. For example, previous studies have examined reductions in private self-awareness among soccer fans, measuring self-concept access via self-report (e.g., "If my team scores a goal I really lose myself completely"; Van Hiel et al., 2007). Of course, the utility of this kind of self-report dependent measure hinges on participants' ability to reflect explicitly, retrospectively, and accurately on their own reduced…self-reflection.

To circumvent this challenge, we designed an fMRI experiment in which participants performed a competitive task both alone and as part of a group; we indexed the salience of participants' own moral norms during competition unobtrusively (i.e., activation in a region of the medial prefrontal cortex (mPFC) identified by an independent self-reference task); and we assessed effects on subsequent behavior using a novel index of participants' willingness to harm competitors versus teammates (assigned in the lab; Cikara, Jenkins, Dufour, & Saxe, 2014). Consistent with previous research, participants harmed competitors more than teammates:

specifically, they selected relatively less flattering photographs of their competitors for public distribution. More critically, the degree to which participants were willing to carry out such harm was associated with the degree to which they exhibited reduced mPFC activation in response to first person moral statements while competing in a team context (but not when competing alone). These results suggest that acting as part of a competitive collective can reduce the salience of one's own moral standards and, in turn, enable out-group harm.

While value-driven beliefs and strategic considerations play a demonstrable role in predicting aggression in inter-coalitional contexts, peoples' behaviors are guided also in large part by how they *feel*. The intergroup literature is replete with research on prejudice as a central explanation, typically measured as an attitude. However attitudes—which in their most basic form are conceptualized as a single dimension of valence, ranging from negative to positive—are often not sufficiently specific to predict behaviors. For example, when do negative attitudes predict neglect, as opposed to fear, or attack (Cuddy et al., 2007)? To better predict which behaviors may arise between coalitions one may be better served by turning to an analysis of discrete intergroup emotions (e.g., Neuberg & Cottrell, 2006; Mackie & Smith, 2015; Stürmer, Snyder, & Omoto, 2005).

3.2 Emotions

Similar to goals and priorities, people's *emotional* responses may shift in intergroup contexts to reflect the interests of the coalition instead of the individual (Mackie & Smith, 2015). Nowhere is this pattern more apparent than in the domain of how people feel in response to the suffering of those within versus without their coalitions.

By some accounts the capacity for empathy is one of the most important faculties that humans possess. Empathy is a powerful motivator of cooperation and altruism, cornerstones of humanity's social uniqueness and success (Batson, 2009; Keltner, 2009; Tomasello, 2009). It comprises both cognitive and affective components that allow people to share, understand, and respond to others' experiences and feelings (Weisz & Cikara, 2020). Empathy is also, for better or for worse, bounded: people do not empathize with everyone, in equal measure, all the time. Specifically, people often report feeling less empathy for individuals who do not belong to their coalitions or categories relative to those that do (Batson & Ahmad, 2009; Cikara, Bruneau, & Saxe, 2011). Consistent with these self-report findings, dozens of neuroimaging and EEG studies report that people show decreased and sometimes absent physiological responses associated with empathy when witnessing out-group relative to in-group members in physical or emotional pain (see Han, 2018 for a recent review). This *intergroup empathy bias* matters because the absence of empathy reflects a reduction in motivation to engage in pro-social behavior toward those who are suffering. Said another way, a lack of empathy for "them" places those people beyond the spheres of morality and justice that we believe apply to "us."

One key insight that I have sought to emphasize is that the absence of empathy is not antipathy; it is apathy or indifference (Cikara & Fiske, 2013; 2014). Apathy is generally not a strong motivator of behavior, perhaps with the exception of neglect. For example, people may cross the street to avoid speaking to an unhomed person, but most of us would be surprised if they were to go out of their way to harass that person. This absence of a relationship between empathy and harm is borne out in a meta-analysis: across 106 effect sizes, empathy and aggression are correlated only $r = -.09$ (Vachon, Lynam, & Johnson, 2014).

Much more than the absence of empathy I have focused on identifying the conditions under which people experience the exact opposite of empathy in response to outsiders' good and

bad fortunes. By contrast to empathy, I have found pleasure in response to others' mis-fortunes — Schadenfreude — or displeasure in response to others' triumphs — Glückschmerz — are feasible motivators of inter-collective conflict and violence (Cikara, 2015). So what predicts which emotions we experience—empathy, apathy, or Schadenfreude—when we see or learn of another person's suffering? While several conditions predict the experience of Schadenfreude in interpersonal contexts (see Smith et al., 2009; Van Dijk, Ouwerkerk, Smith, & Cikara, 2015 for reviews), here, I will focus on the effect of inter-coalition competition. Note however that the example I just cited—of feeling indifferent toward rather than pleased about the suffering of an un-homed individual—would suggest that competition alone is insufficient. Therefore we predicted that the specific combination of perceptions of a collective's competitiveness and status—their desire and ability to enact a threat—would be most likely to elicit counter-empathic emotions (Cikara & Fiske, 2013; 2014).

In order for Schadenfreude to qualify as an intergroup emotion, people must feel it on behalf of their group; however, people only appraise events from an intergroup perspective when they are highly identified with the relevant in-group (Mackie, Devos, & Smith, 2000). Another constraint one has to consider is that Schadenfreude is a socially undesirable emotion (Smith et al., 2009). As such people may be somewhat reticent to report it. However, sports, politics, and celebrity gossip are a few domains in which it is acceptable, even desirable, to express pleasure at others' misfortunes, and so sports teams have become the fruit fly of intergroup Schadenfreude research. The empirical evidence bolsters the notion that identification is a true constraint on intergroup Schadenfreude. For example, college basketball fans' identification with their team predicted greater Schadenfreude in response to a rival player's injury. Fans' Schadenfreude, in turn, correlated with greater disappointment in response to news that the injury did not end the rival player's season (Hoogland et al., 2015). In another example, soccer fans smiled more intensely, as measured by facial electromyography (EMG), when they watched a rival soccer team miss a penalty kick relative to when they watched their favored team make the goal (Boecker, Likowski, Pauli, & Weyers, 2014). Of course, Schadenfreude and Glückschmerz are natural responses in zero-sum contexts; if 'they' are unhappy, 'we' are pleased. This requires that intergroup Schadenfreude experiments include pure 'spite' conditions—those in which the threatening coalition suffers without any tangible benefit to one's own coalition.

The first experiment we ran to probe the relationships among perceived coalitional competitiveness and status, Schadenfreude, and harm was in the context of a real-world conflict: Red Sox versus Yankees fans, historic rivals in American baseball (Cikara, Botvinick, & Fiske, 2011). We pre-screened our participants for hardcore fandom: they had to love their own team and dislike the rival team; had to identify correctly players from photos; and had to know specific players' positions. As predicted, pre-experiment survey data with our sample confirmed that all our participants rated their favored team as most warm (an attribute associated with absence of competitiveness) and competent (an attribute associated with status), the Orioles (a relatively less competitive team in the same league) as moderately warm and moderately competent, and most important, their rival as admittedly more competent than the Orioles, but also less warm.

During the main experiment, Red Sox and Yankees fans underwent fMRI while viewing animated baseball plays involving their favored team, rival team, and two other teams (the Orioles and the Blue Jays, two non-rivals in the same league), succeeding and failing to get on base or getting tagged out. Following each play, participants reported how much pleasure, pain,

and anger they experienced watching the play unfold. Unsurprising, participants said they felt the most pleasure and the least pain and anger when their favored team scored against their rival, their rival failed to score against their favored team, and critically, their rival failed to score against the Orioles as compared to plays in the control condition (the Orioles failing and succeeding against the Blue Jays). This last condition in which the rival failed against the Orioles was the pure Schadenfreude condition because the favored team did not benefit in these cases; the pleasure came only from spite toward the rival. In addition, participants said they felt more anger and pain when their favored team failed to score against their rival and their rival scored against their favored team as compared to the control condition. Finally, one to two weeks after participants had been scanned, they completed a follow-up survey in which we asked them how likely they would be to enact a variety of hostile behaviors both toward rival fans and Orioles fans. Both Red Sox and Yankees fans reported that they were more likely to heckle, insult, threaten, and hit a rival fan as compared to an Oriole's fan.

We analyzed our fMRI data to test whether empathy and pleasure in response to out-group pain relied on separable neural circuitry. A wealth of existing research had already established a strong correlation between dorsal anterior cingulate (dACC)/anterior insula (AI) responses and self-reported empathy (see e.g., Decety, 2011); the link between ventral striatum (VS) and reward (specifically, reward prediction error) was even better established and conserved across numerous species (see e.g., Bartra, McGuire, & Kable, 2013). As predicted, painful baseball plays increased responses in ACC and AI. By contrast, pleasurable baseball plays, including rivals failing to score against the Orioles (the pure Schadenfreude condition), increased responses in the VS, the region associated with learning from unexpected *rewarding* events. Weeks later, those participants who exhibited greater VS activation in response to watching their rivals fail also reported an increased likelihood of aggressing against rival team fans (relative to Orioles fans). In fact, VS activation in response to watching rivals fail was a better predictor of harm than even participants' subjective reports of pleasure in response to watching rivals fail. Note also that no such correlation emerged with dACC or AI (mirroring the absence of a relationship between reduced empathy and aggression). However these findings are based on a very small sample (n = 18) and so should be interpreted with caution.

That said, our findings dovetailed nicely with another fMRI study of intergroup Schadenfreude that was published at the same time: only this one was with soccer fans (Hein, Silani, Preuschoff, Batson, & Singer, 2010). In the first phase of the experiment soccer fans received or witnessed fellow and rival fans receive electric shocks. In the second phase participants only witnessed fellow and rival fans receive electric shocks, but had an opportunity to volunteer to absorb some of the shocks themselves to reduce pain to others. Those participants who had more negative views of the rival exhibited more VS when witnessing rival fans receive electric shocks in Phase 1. More important, those same participants who exhibited greater VS in response to rival pain also provided the least aid to rival fans in Phase 2. Thus the intergroup Schadenfreude/VS relationship emerges in response to actual physical pain (not just team outcomes in some abstract sense) and occurs in the case where targets are merely affiliated with the team (fans, not players themselves). Together, these two experiments were the first to establish a link between intergroup Schadenfreude and endorsement of harm against (and withholding of help from) a competitive collective and its members. It is worth noting that the VS/Schadenfreude relationship replicates in several interpersonal fMRI studies as well, but only when people are in competition with one another or retaliating for past harm (de Bruijn, de

Lange, von Cramon & Ullsperger 2009; Chester & DeWall, 2016; Singer, Seymour, O'Doherty, et al., 2006; Takahashi et al., 2009).

Of course it is part of the script of sports rivalries that people are allowed, even encouraged to express emotions like Schadenfreude. Thus our next aim was to test whether people experience Schadenfreude in more subtle contexts: when targets of misfortunes are merely stereotyped as competitive and high-status (e.g., Asians, female professionals, investment bankers). In other words, can observers can experience Schadenfreude in the absence of any interaction (e.g., without history of conflict, explicit competition)? Again, this is of interest to me because these emotions may facilitate tolerance, or even commission of harm. A complementary theoretical implication is that these fundamental coalitional cues—status and competition—enable us to generalize our predictions and results to a wide variety of social categories and contexts.

Why might encountering members of categories who are merely stereotyped as competitive and high-status be sufficient to engender Schadenfreude when they experience misfortunes? Social comparisons happen automatically (Wedell, 1994). As such, simply encountering a target whose category is stereotyped as high-status may make one's comparatively lesser status more salient than at baseline. If this target's category is *also* stereotyped as competitive, social emotion theory predicts an observer will be likely to experience a contrastive emotions, such as envy (rather than assimilative emotions like admiration; Smith, 2000). Envy is a strong predictor of Schadenfreude in interpersonal contexts (Smith et al., 1996) thus we predicted that envious prejudice (Fiske et al., 2002; 2007) would be a strong predictor of Schadenfreude at the collective level of analysis (Harris, Cikara, & Fiske, 2008). Again, feeling pleasure instead of empathy disrupts the link between observing others' suffering and being motivated to help them. Therefore we finally predicted that categories that are most likely to elicit Schadenfreude would also be more likely to be subject to harm.

In our first experiment testing these hypotheses, participants viewed a series of positive, negative, and neutral events, each paired with an unlabeled photograph of an individual on a white background (e.g., a drug addict, an elderly woman, a man in a business suit; Cikara & Fiske, 2012). Critically we included multiple categories for each of the four quadrants made up of the competition-by-status features; this ensured that we were testing the effects of the features rather than the categories themselves. Each image had been validated in a separate sample as evoking the critical stereotypic traits (i.e., the correct levels of warmth and competence—attributes associated with competition and status, respectively).We also made sure the events were misfortune that could befall anyone: "Ate a really good sandwich," "Got soaked by a taxi," and "Yawned twice in a row." Also important: none of the events described the target taking an action for which they were being rewarded or punished. We did this to eliminate deservingness as a confound (i.e., to avoid a reaction such as, "I'm pleased this person suffered a misfortune because they deserved their comeuppance").

After each trial, we instructed participants to answer two questions: "How GOOD [BAD] would this make you feel?" We specified that "this" referred to the target's experience described on the previous screen. Note that we asked participants these two questions separately to allow for the expression of ambivalence (e.g., they could say they felt simultaneously bad *and* good). One limitation is that asking how good and bad individuals feel in response to positive and negative events cannot fully capture the constructs "empathy" and "counter-empathy." Though we did not have this data at the time we ran the stereotype study, a subsequent pilot study (N=353; Hudson, Cikara, & Sidanius, 2019) assessed negative empathy and Schadenfreude

using a multi-item scale that included the "good" and "bad" items as well as how sad/sympathetic/compassionate/concerned the participant felt (for empathy; items from Stürmer, Snyder, Kropp, & Siem, 2006) and how relieved/happy/satisfied the participant felt (for Schadenfreude; items from Leach et al., 2003;). Each subscale was internally consistent: empathy alpha = 0.87; Schadenfreude alpha = 0.96. Furthermore, an exploratory factor analysis indicated that a 2-factor solution that separated empathy and Schadenfreude best characterized the data. Finally, and most important, the "how bad [good] does this make you feel" items had some of the highest factor loadings on the empathy and Schadenfreude factors indicating that these items in isolation are valid measures of empathy and Schadenfreude.

There was, however, another challenge. The existing literature suggested that participants rarely self-report feeling Schadenfreude in excess of the midpoint of a scale (e.g., van Dijk, Ouwerkerk, Goslinga & Nieweg, 2005, Smith et al., 1996; Leach & Spears, 2008, 2009) indicating that participant responses were potentially distorted by social desirability. To assess another, less controllable indicator of affect, we also recorded participants' facial muscle movements using EMG. We predicted that participants would exhibit the most positive affect (i.e., smiling), and not just reduced negative affect, in response to stereotypically competitive/high-status targets' misfortunes. We focused specifically on the ZM (zygomaticus major; a cheek muscle) because it is the muscle responsible for pulling the corners of the lips into a smile and correlates reliably with positive affect (Brown & Schwartz, 1980).

In a second phase of the experiment we presented each of the targets from the first phase, this time without any events, and asked participants to rate "As viewed by society, how COMPETENT [WARM] is this person?" This merely served as a manipulation check to confirm that our participants' stereotype assessments of each target's warmth and competence replicated the sample on which the stimuli were normed—in short, they did.

As predicted by a coalitional account, participants self-reported that they felt the least bad about negative events, and least good about positive events for stereotypically competitive/high-status targets. In line with our concerns about social desirability, participants did *not* report feeling significantly better in response to negative events for stereotypically competitive/high-status targets compared to the other targets. By contrast, the facial EMG measures supported all of our hypotheses (Figure 4). Participants smiled more in response to *negative* relative to positive events (an incongruent, Schadenfreude response) when they were paired with stereotypically competitive/high-status targets. This indicates the presence of *positive affect*, not just the absence of negative affect, in response to competitive/high-status targets' misfortunes. For all other targets, participants smiled more in response to positive relative to negative events.

Our EMG study provided us with correlational evidence that stereotypes comprising high-status and competitiveness were sufficient to evoke Schadenfreude; however, we wanted to test the causal influence of these coalitional features (and eliminate the possibility that there were other features of the targets driving our results). We hypothesized that increasing the perceived cooperation and decreasing the status (versus not) of the same stereotypically competitive/high-status target should reduce observers' Schadenfreude. Therefore, in another experiment we manipulated the competitiveness and status of one particularly competitive, high-status category: investment bankers (Cikara & Fiske, 2012). Participants read a newspaper article about one of the following: investment bankers whose situations were status quo (competitive/high-status)); i-bankers who were advising small businesses, pro bono, to help the economy as a whole (decreases competitiveness); i-bankers who were using the last of their bonuses to fund their drug habits (decreases status); or i-bankers who were unemployed but still dressing up in their

suits and pretending to go to work (decreases both status and competitiveness). We predicted that participants would report feeling significantly worse about investment bankers' misfortunes after reading any of the vignettes relative to the status quo vignette. As predicted, decreasing status and competition information about investment bankers increased participants' empathy, but only for those targets who resembled investment bankers (not other competitive/high-status targets). Thus, this study established more clearly that these coalitional features are driving Schadenfreude responses.

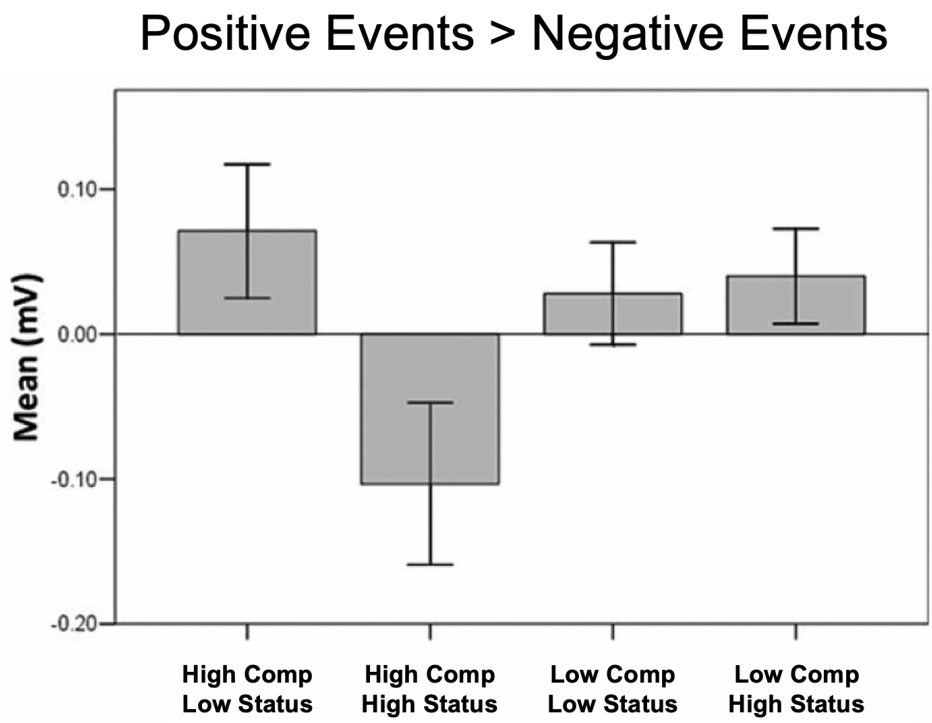

Figure 4. ZM response during negative events minus ZM response during positive events—only competitive/high-status targets elicited more ZM response during negative as compared to positive events. Bars represent SE. Adapted from Cikara & Fiske (2012).

What about willingness to harm? In a follow up fMRI study we assessed self-reported affect as well as participants' willingness to harm a variety of targets who varied on competitiveness and status (Cikara & Fiske, 2011). While they were in the scanner, participants viewed the same target-event pairs from the EMG study though they did not report how they felt until after the scan when they viewed target-event pairs again and reported their affect on a bi-polar scale (1 extremely bad to 9 extremely good). Replicating the results of the EMG study, participants reported feeling worst about positive events and best about negative events when they were paired with competitive/high-status targets as compared to all other targets. About two weeks after the scanning session we contacted participants with a follow-up survey, which presented the following scenario: "You are participating in a Fear Factor type game show and have just won a challenge. This exempts you from the 'punishment' the rest of the players face: they are all going to receive mild electric shocks, which are painful, but not lethal. The game show host gives YOU the choice to decide whether all five of the players are going to get shocks or if one person should get a stronger shock (which is again, painful, but not lethal) while you

spare the other four." We then asked how willing they would be to volunteer each person they saw during the scan to receive a shock so that the other, unidentified players could avoid the pain. As predicted, participants were significantly more likely to subject competitive/high-status targets as compared to all other targets to painful electric shocks.

One notable limitation of these previous studies is that our previous competition and status manipulations—sports rivalries and stereotypes—provide more than just competition and status-relevant information. What happens when we strip out all social information except that there are two collectives who are or are not in competition with one another? In the next series of experiments in this line of work, we manipulated functional relations between *novel* groups, with no history of conflict and no stereotypes associated with them, to determine whether mere inter-group competition was sufficient to alter participants' empathy toward own team and other-team members experiencing good and bad fortunes (Cikara, Bruneau, Van Bavel, Saxe, 2014). In each experiment, we randomly assigned people to either the Eagles or the Rattlers team, ostensibly based on their personality profiles. We told them that the two teams were involved in an ongoing problem solving challenge and the two teams were neck and neck (though the other team was just slightly ahead; this was held constant across all conditions to signal that they were a capable team). In the competitive condition, whichever team reached 100 points first would receive a monetary bonus and the other team would receive nothing beyond their base pay; in the independent/neutral condition each team would receive a monetary bonus when they collectively reached 100 points total, irrespective of what had occurred with the other team; in the cooperative condition we said that the teams should work together to reach 200 points and only then would everyone receive a monetary bonus. In the next phase we told them that scientific evidence suggests that people perform better in these problem solving challenges when they know something about the other players. To that end, they would read about recent experiences of same- and other-team players which had been shared with us by past participants (though we had standardized the language). Following that, participants saw a series of good and bad fortunes associated with Eagles and Rattlers players. On each trial, participants reported how good and how bad they felt. Critically, the events (e.g., "accidentally stepped in dog poo") had no bearing on each target's ability to partake in the problem-solving challenge. As such, there was no strategic benefit conferred by other-team members' misfortunes.

As predicted, in competitive contexts, participants reported experiencing more empathy for same-team than other-team targets and more counter-empathy for other-team than same-team targets (Figure 5). This effect was attenuated in the independent condition and entirely absent in the cooperative condition. In a follow-up control experiment, we confirmed that our effect was not driven by participants' bias in their perceptions of how *targets* felt in response to positive and negative events; they were perfectly happy to report that competitors felt as bad as teammates when bad things happened to them.

Of course any time one documents a bias in feelings toward one versus another collective it remains ambiguous whence the effect comes: specifically in this case, extraordinary empathy for one's own team or animus for the other team? In the next experiment we restricted the design to include only the competitive structure, however we added a third collective. We told participants that some people did not fit the profile of either the Rattlers or the Eagles, but that we did not want to exclude those people from participating and earning money. Their stories would appear interspersed among Rattlers' and Eagles' stories but the page would be blank where a team logo would otherwise appear. We had two competing predictions. If participants reported similar levels of (counter-)empathy for unaffiliated targets and competitors then we

could infer that our previous results were driven primarily by increased empathy for one's own team. If, however, participants reported similar levels of (counter-)empathy for unaffiliated targets and teammates, then we could infer that our results were driven more by out-group hostility. Consistent with the second hypothesis, unaffiliated targets received empathic responses indistinguishable from teammates, demonstrating that intergroup empathy bias in competitive contexts is better characterized as competitor apathy/antipathy than extraordinary teammate empathy.

Figure 5. Empathy (upper left, lower right) and counter-empathy (upper right, lower left) ratings for novel in-group and out-group targets under competitive, cooperative, or independent functional relations. Plotted values are least squares mean estimates and standard errors computed from the omnibus model. Adapted from Cikara et al. (2014).

One important next step for this research program is to elucidate the temporal dynamics of the relationship between Schadenfreude and harm. I have proposed that the capacity for collective violence may have developed, in part, by appropriating basic reinforcement-learning processes and associated neural circuitry in order to overcome harm aversion. That is, the cycle of intergroup aggression may begin with the passive observation of a negative outcome for another person. Which emotion an observer experiences will depend on the target's coalition membership. If the target is a competitive coalition member, the observer may be surprised to find they experience some Schadenfreude. This pleasure may potentiate a small aggressive action (or at least the withholding of help), the outcome of which is the continued or increased

suffering of the target. Assuming the target's identity is still "threatening coalition member," this should be further cause for pleasure, invigorating further aggressive behavior.

What evidence is there that Schadenfreude may contribute to learning to overcome harm aversion? Neuroscience plays a key role here because it allows us to leverage what we have learned from decades of research on the biological bases of reinforcement learning. There are several regions of the brain that support encoding and representing subjective value, but VS supports a critical component of reinforcement learning in particular—encoding rewarding events for the purposes of learning which actions will increase the likelihood of reward in the future (Bartra, McGuire, & Kable, 2013; O'Doherty, 2004). The consistent relationship between Schadenfreude and VS engagement implicates not only the VS's valuation function (i.e., evaluating competitor harm as positive), but also its motivation function (i.e., biasing action selection toward behaviors that harm competitive coalitions and associated individuals). This suggests that the repeated experience of pleasure in response to competitive coalition members' suffering may make people more likely to become first person agents of harm when given the opportunity. As we have already reviewed, VS activation increases in response to rivals' suffering (e.g., in the context of the baseball and soccer studies described above) and is associated with an increased subsequent desire to harm (Cikara et al., 2011) and decreased willingness to help rival fans (Hein et al., 2010). Interesting, even rats exhibit a dopamine spike within VS after attacking an intruder (Van Erp & Miczek, 2000); mice also exert effort to harm other submissive mice (Legrand, 2013). Thus there is evidence across species to suggest that in some contexts harm carries intrinsic value just like other reinforcers (Chester, 2017).

We have some preliminary evidence to suggest that people who are particularly hostile toward an out-group experience out-group directed spite as a reinforcer (Moore et al., in preparation). Across three behavioral studies, we had participants complete an incentivized probabilistic decision-making task in which they could learn via trial-and-error to obtain individual or own-team benefits (which had no effect on the other team) or those same benefits while also causing the other-team harm (what we call "spiteful" actions); both outcomes were associated with equivalent monetary bonuses (see Figure 6). Unlike other related tasks, each action in the task was probabilistically associated with a helpful versus spiteful outcome, which meant that even the most harm-avoidant individuals sometimes took an action that resulted in harm, giving them a chance to update their preferences via experience. We then used a reinforcement-learning approach to model participants' latent preferences for pure own-team help versus other-team spite.

Critically, we structured the task so that it was costly to exhibit a preference both for and against other-team harm. Our goal was to attenuate socially desirable responding (i.e., avoiding spiteful behavior) or responding in line with a demand characteristic (i.e., enacting an intergroup script by pursuing other-team harm in excess of one's personal preferences). Specifically, in our task, money-maximizing players ought to have been indifferent between actions associated with a lower versus higher likelihood of other-team harm because both outcomes yielded the same amount of money to the self when they paid out. However, a preference for either avoiding or seeking other-team harm would bias participants to select actions associated with either low or high probability of harm, even during low payout periods for those actions. This is how exhibiting a preference became costly: participants gave up real money to take subjectively preferable actions. Finally, we examined whether subjective valuation of other-team spite was correlated with negative out-group attitudes to bolster our confidence that participants' behavior reflected their social preferences.

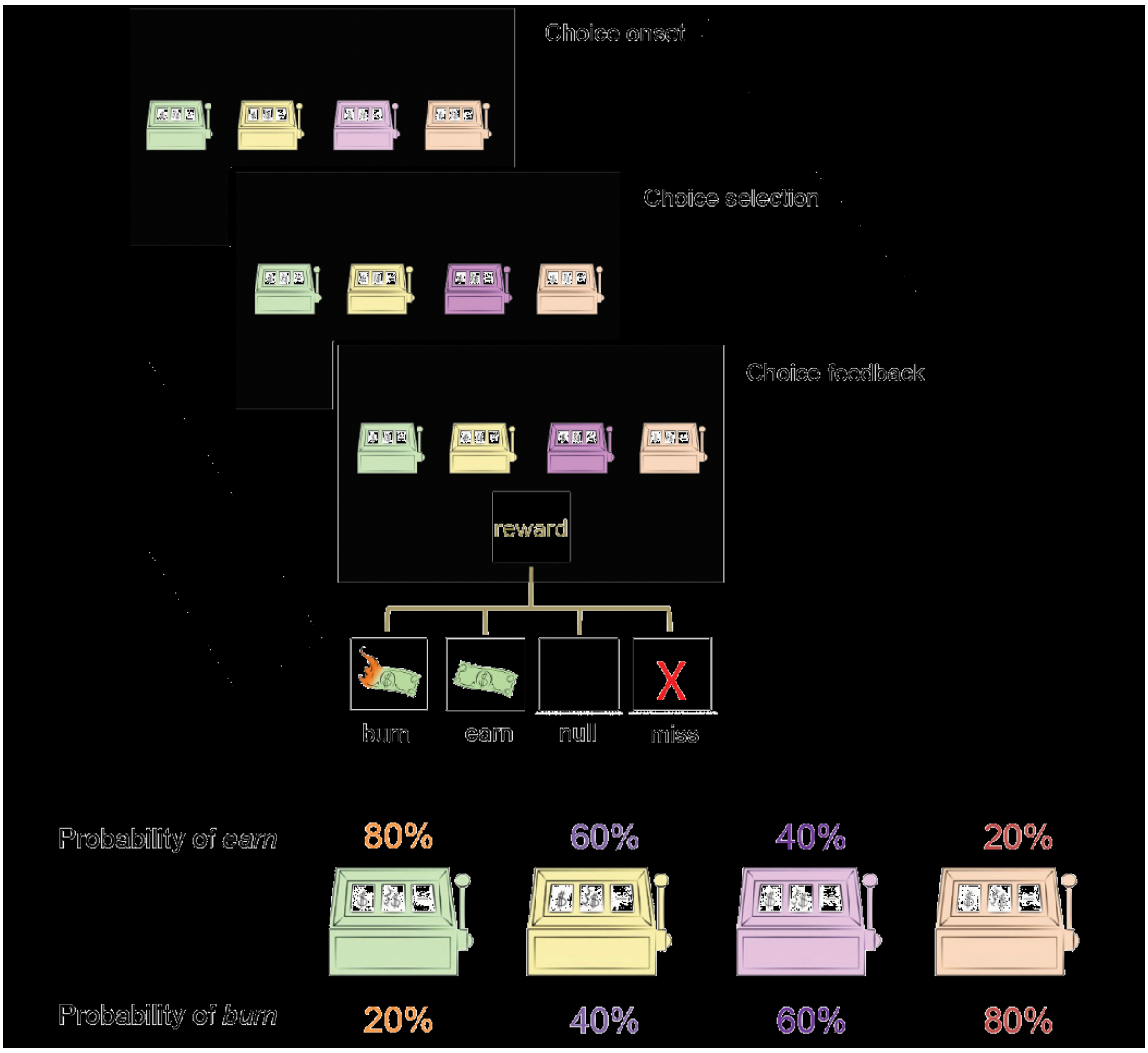

Figure 6. Schematic of task from Studies 2 and 3 (Moore et al., in preparation).Top panel: On each trial participants selected one of four slot machines and then received feedback. Each machine could yield a burn (earn money for self/own-team *and* subtract money from competitor), an earn (earn money for self/own-team), or a "no points" outcome ("miss" was only an option in the fMRI version of the task if participants did not answer within the trial time). Bottom panel: Each machine paid out with a drifting probability (orthogonal to one another), however each machine, when it *did* give a reward, had a fixed probability of yielding an earn versus a burn. As such, participants could learn via experience which machines were burn-biased and which machines were earn-biased. Adapted from Ichikawa, Moore, & Cikara (2019).

Across these three studies, and in contrast to previous research (albeit with very different paradigms that did not allow for learning via experience; e.g., Halevy, Bornstein, & Sagiv, 2008), participants exhibited a small (or absent) aversion to spite. At the individual level, more negative other-team attitudes correlated with a greater preference for spiteful actions. These effects held across hypothetical and political group conflicts, and both when participants were

benefiting themselves and their coalitions. In short, greater dislike of the other team was associated with greater subjective valuation of spiteful out-group harm. These results may have important implications for escalation of inter-coalition aggression. If harm for harm's sake acts as a reinforcer for some, those people may be more likely to engage in harm without provocation. These *proactive* aggressors are dangerous because their actions may incite other, less inclined individuals to aggress as well; for example, people are twice as likely to imitate aggressive behavior when it targets an ethnic out-group member relative to coethnics (Bauer, Cahlíková, Chytilová, & Želinský, 2018). These and our findings together suggest that it may take only a few proactive individuals to generate a contagious spiral of collective violence.

Another important future direction for this work is to determine how these emotions serve to motivate behaviors that reinforce individuals' coalition-related ideologies. We have begun to examine this question specifically in the context of Social Dominance Orientation (SDO; Hudson, Cikara, & Sidanius, 2019). SDO is an ideological variable indexing how much people prefer and promote group-based inequalities. It is an excellent candidate as a moderator of intergroup emotions because it is associated with a competitive worldview in which resources are zero-sum and some groups deserve more than others. To the extent that people are predisposed to view everything as a competition, they should also exhibit reduced empathy and increased counter-empathy toward everyone, but especially toward those whose oppression maintains the social hierarchy. Across three studies we found that higher SDO scores among White participants were associated with less empathy and more Schadenfreude in response to others' bad fortunes—this was for people in general. However, when we primed White respondents with symbolic threat, they reported significantly less empathy and more Schadenfreude for Asian and Black targets than for White targets. In a final study we found that this pattern replicated even in novel groups, so long as the groups were competitive: higher SDO scores were associated with decreased empathy and increased counter-empathy for competitive team members relative to one's own team members.

Do these emotion patterns facilitate hierarchy-reinforcing behaviors and lead people to avoid those behaviors that might undermine the hierarchy? If so, people with differing ideologies should be *motivated* to experience or avoid empathy and Schadenfreude. In ongoing work we've found that people will relatively higher levels of SDO not only *desire* to feel less empathy and Schadenfreude toward low-status targets, but when given a choice, *choose* to feel less empathy and more Schadenfreude (Hudson, Cikara, & Sidanius, under review).

Thus, consistent with a coalitional account, high-status, competitive collectives are more likely to be targets of Schadenfreude and harm than other collectives who do not exhibit those features. The novel groups results illustrate the fluid nature of this phenomenon: groups need not have a long history of interaction to elicit malevolent affective reactions. This suite of findings are promising for at least two reasons. First, focusing on generalized features such as perceptions of status and competitiveness affords us the ability to make predictions about when and which collectives will be at greatest risk, particular in times of social instability when threat is heightened for all collectives in an environment. Second, knowing that these perceptions are malleable or context-bound, rather than "essential" to groups themselves, means it is possible to combat hostile emotional and behavioral responses. Giving people this knowledge may empower them to second guess their more cruel impulses across a variety of ethically consequential contexts including, but not limited to policy preferences, discrimination, and inter-coalitional conflict.

3.3 Attributions of emotion

Just as cues to coalition competition and status shape our own emotions, so too do they shape our attributions and expectations of emotion to other-coalition members. Accurate expectations of emotions matter in inter-coalitional contexts because our perceptions and forecasts of the other side's emotions inform our decisions about and behaviors toward them.

In the first series of studies we ran examining attribution of emotion we focused on judgments of facial displays of emotion (Lazerus et al., 2016). We tested two competing hypotheses: an own-team accuracy hypothesis (that people would be more accurate judging own-team emotion displays) and an own-team positivity bias hypothesis (that people would simply judge own-team emotion displays as more positive than competitors', irrespective of which emotion they were displaying). In the first experiment we randomly assigned people into two novel, competitive groups—the Green Team and the Blue Team—who were competing for a bonus. Participants then rated the valence of both same-team and competitor-team members' fearful, happy, and neutral facial expressions using a two dimensional grid including valence (negative to positive) and arousal (low to high) dimensions. Irrespective of which emotions targets displayed, participants judged own-team members' expressions as more positive than competitor-team members' expressions. In the next experiment participants categorized same-team and competitor-team members' fearful and happy expressions as either "positive" or "negative" in a mouse-tracking paradigm. Participants exhibited the most direct trajectories toward the "positive" label for same-team happy expressions but also an initial attraction toward "positive" for same-team fear faces. The competitor team trajectories were intermediate between the two same-team trajectory extremes. Finally the last experiment replicated Experiment 2 and demonstrated that the effect held irrespective of whether targets' gaze was direct or averted. In summary, we found support for the own-team positivity bias: people judged same-team faces as more positive than competitor-faces, regardless of emotion, across multiple modes of measurement.

In the second line of experiments in this emotion attribution line of research we moved away from facial displays to examine affective forecasts for same-category, competitive-category, and unspecified people in intergroup contexts (e.g., elections, football games; Lau, Morewedge, & Cikara, 2016). Typically, providing forecasters with more relevant information improves accuracy (Gilbert & Wilson, 2007). By contrast, we found that providing forecasters with targets' social-category information made forecasts more extreme and less accurate. In both political and sports contexts, forecasters across five experiments overestimated the emotional reactions of hypothetical targets when targets were labeled by category membership relative to when they were unspecified. For example, when asked how people in the U.S. would feel if their party lost/failed to gain the majority in the 2014 Senate midterm election, participants said targets would be significantly unhappier (when the target was labeled as a "Democrat" or "Republican" relative to just "person." Importantly, this made forecasters less accurate: forecasts for category labeled targets were more extreme than experiencers ratings of how they *actually* felt. This overestimation effect held for both same-category and competitive-category members and when predicting responses to positive and negative outcomes.

What was the source of this forecasting in accuracy? Having participants make judgments under time pressure reduced the extremity of forecasts for category-labeled but not unspecified targets, suggesting that the effect was due to overcorrection for social-category information rather than differing priors for category-labeled versus unlabeled targets. What we discovered was that stereotypes, rather than retrieval of extreme exemplars, better accounted for this

overcorrection. Specifically, participants forecasted that a stereotypically unreactive but still labeled category—Buddhists—would be least unhappy after suffering a loss in an online game, followed by unspecified targets, with political party-labeled targets rated as most unhappy (Figure 7). However, we also had participants rate the relative extremity of the person that came to mind from each category. After making their forecast, participants indicated how that target would rank among 100 of his or her peers who also watched their group lose the tournament. Forecasters did not report recruiting more extreme exemplars when making predictions for category-labeled targets. Therefore, stereotypes, in particular those associated with competitive categories, drove overcorrection, making forecasts less accurate.

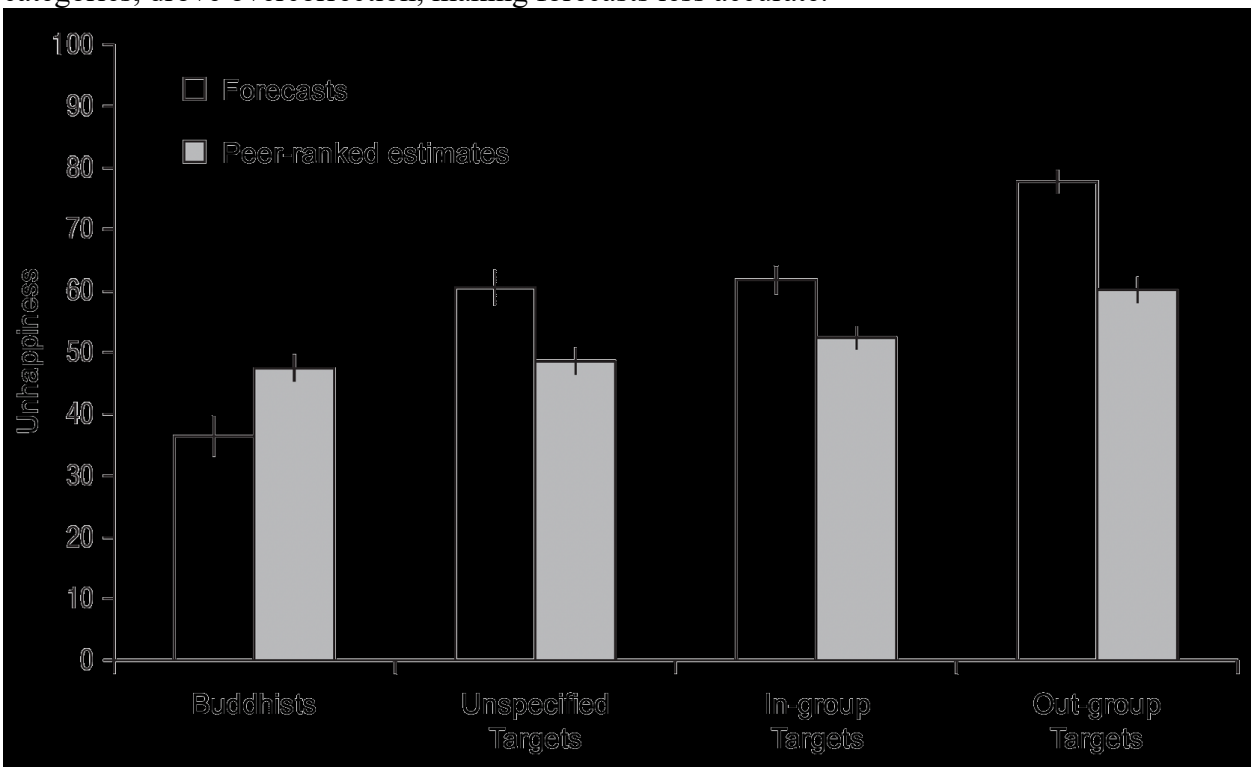

Figure 7. Mean forecaster rating of how unhappy each of the four targets would feel if their team lost, along with peer-rank estimates for each target. On the y-axis, 0 indicates neutral affect. Error bars represent +/- 1 SEM. Adapted from Lau et al. (2016).

Nowhere is the importance of accurate affective forecasts more consequential for inter-coalitional behavior than group-meta perceptions. What ‘we’ believe ‘they’ feel about us and our behavior contributes to our assessments of the possibility of inter-coalitional coordination versus conflict. Inaccuracy, more specifically pessimism, in these estimates (e.g., I believe that you’re outraged by my coalition and everything we do) forecloses on the possibility of working together effectively (Lees & Cikara, 2020). Our first aim was to test whether people were inaccurate in their group meta-perceptions. In one of the experiments we asked a representative sample of Democrats and Republicans to read a series of scenarios about interparty sabotage. The scenarios ranged from one party changing a highway name so it was named for a beloved party member to one party gerrymandering a district in their favor. Each respondent was randomly assigned to one of three treatments: an actual perception condition, an in-group perception condition, or a group meta-perception condition. In the actual perception condition we just asked Democrats and

Republicans: how much would *you* dislike it if the other party engaged in this behavior? In the in-group perception condition we asked people: how much would the *average fellow in-group member* dislike it if the other party engaged in this behavior? Finally in the group meta-perception condition we asked: how much would the *average out-group member* dislike it if YOUR party engaged in this behavior? Averaged across scenarios Democrats and Republicans in the actual perception condition indicated dislike at only about the midpoint of the scale: 55 out of 100. What about estimates for a fellow party member? Here we already began to see inaccuracy. People reported that fellow party members would be significantly more upset—about 11 points more—than Democrats and Republicans actually were. Finally, in the group meta-perception condition, the average dislike rating was 77 with many responses clustering at 100: maximal dislike. In other words people's meta-perceptions of how upset the other side would be were inflated by 40% relative to how the other side said they would actually feel. Democrats and Republicans exhibited this inaccuracy in equal measure and the pattern of data was identical when we asked how opposed respondents were to the behavior and how unacceptable they thought the behaviors were. Note that when we changed the behaviors so that they were cooperative (e.g., gerrymandered a distract in the *other* party's favor) group meta-perceptions were accurate. Thus group meta-perception inaccuracy is not just an extremity bias in how we believe the out-group will react to the in-group's actions in general, but rather, and more specifically, a negativity bias in competitive contexts.

Recall also that in this first study we asked participants to respond to a series of scenarios that varied in how severe they were in undermining democracy. In a re-analysis of our data, we found that respondents were sensitive to severity when making judgments about fellow party-members' reactions but not out-group reactions (Lees & Cikara, in press; Figure 8). For instance, the average Democrat in the fellow in-group conditions reported that they believed other Democrats would be more upset about gerrymandering than renaming a highway. This was not the case for the group meta-perception condition; "upset" ratings were high irrespective of the behavior under consideration. Again, as an example, the average Democrat reported that they believed the average Republican would be equally upset across all the scenarios. This pattern suggests that overly negative group meta-perceptions do not result from a lack of knowledge per se (in this case the relative extremity of the scenarios), but rather from an inability (or unwillingness) to apply the knowledge we already possess.

Are these misperceptions actually related to estimates of the possibility of coordination? In a follow-up study we replicated just the group meta-perception condition and then asked participants to what extent they thought the opposing political party was driven by purposeful obstructionism. As predicted, greater meta-perception inaccuracy was associated with increased belief that the other party is motivated by purposeful obstructionism. However, this was just a correlation study. As such, we followed up with an intervention experiment. Again, we replicated just the group meta-perception condition but then assigned participants to one of three interventions: one that showed them what their meta-perception rating had been on the previous page (control), one that informed them of the other party's average response (meta-perception correction), and one that informed them both of their own party's and the other party's average response (in-group and meta-perception condition). The correction intervention worked; participants who were assigned to the group-meta perception correction condition reported lower ratings of out-group obstructionism than did the control group and the in-group correction conferred no additional benefit.

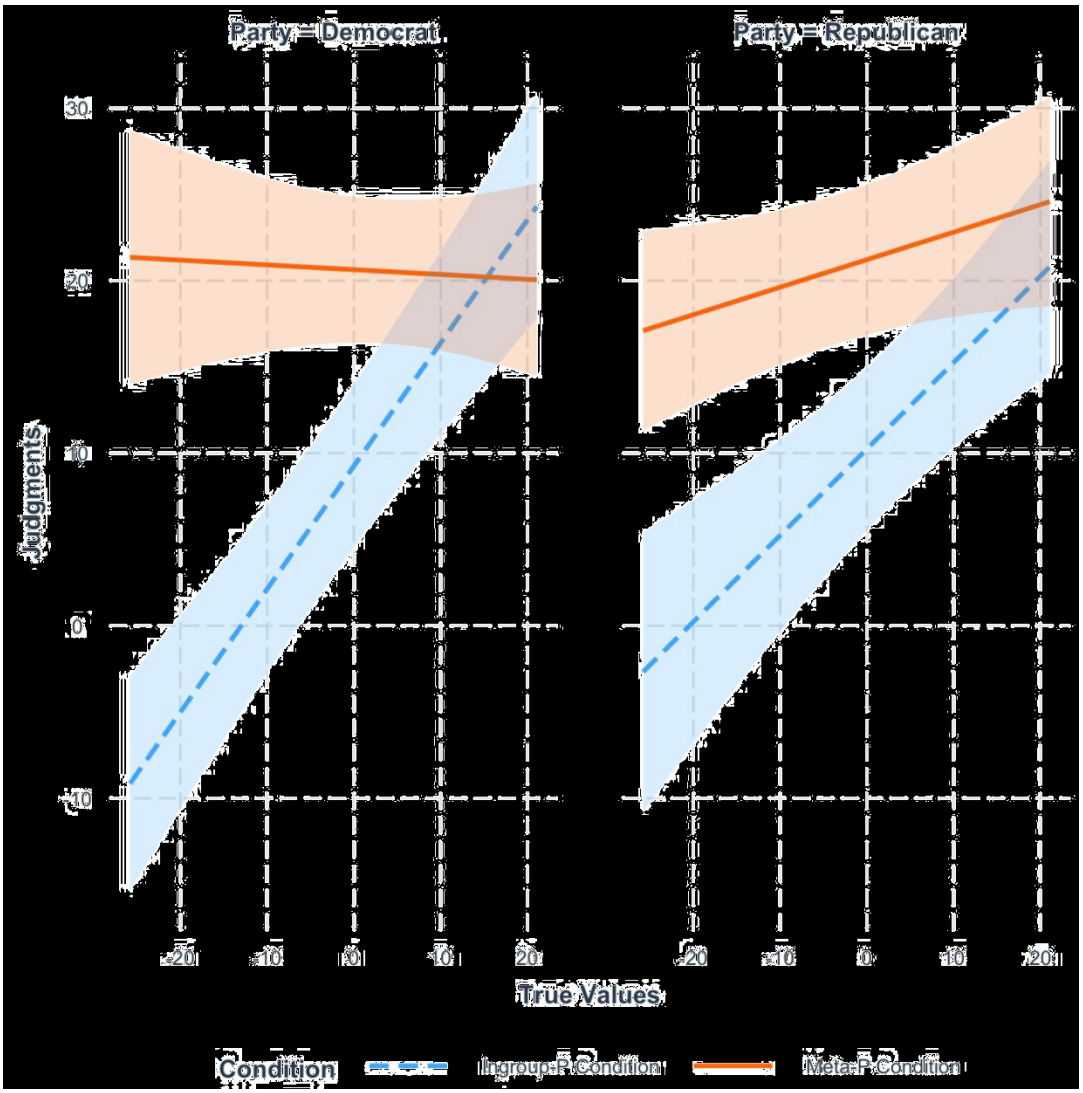

Figure 8. N = 366 (N observations = 5479). Plot of the three-way interaction between party-identification (by panel), condition: in-group perception (blue-dashed) vs. meta-perception (orange solid), and other participants' actual or *true* values in predicting participants' judgments. For both parties, results indicate significant and positive linear relationships between in-group perceptions and corresponding "true" values, providing evidence for rank-related accuracy; there is no such evidence for rank-related accuracy in group meta-perceptions. Error bars 95% CIs. Adapted from Lees & Cikara, (2021).

Particularly exciting is that these results replicate and generalize across cultures (Ruggieri et al., under review). A team of over 80 researchers replicated both our basic meta-perception effect and our intervention in a sample of over 10,000 participants spread over 26 countries (using a variety of competitive social groups). Thus we are reasonably confident that this is not a uniquely U.S. or even political phenomenon.

Unlike intervening on first-order polarization—e.g., trying to change people's actual issue positions or prejudices—combating misperceptions of polarization is just about improving accuracy: specifically, improving accuracy in how 'they' see 'us.' This highlights one way that meta-perceptions are uniquely powerful. They are ultimately about ourselves. The drives to be liked and respected are core social motives (Fiske, 2018b). Thus people may be uniquely

sensitive to corrective information about how others see them and their groups (particularly if it is in a relatively more positive light)—certainly more sensitive than they are about how wrong they are about "them."

Again, consistent with a coalitional account, cues to competitiveness reliably distort emotion attribution to other-coalition members across a wide variety of methodological assessments and intergroup contexts. These distortions have demonstrable implications for predictions regarding inter-coalition coordination and hold some potential for reducing suspicion amongst coalitions in conflict.

4. INTERVENTIONS

One of the primary benefits of a coalitional account is that it highlights the specific cues that drive the cleaving of collectives into "in-groups" and "out-groups," irrespective of whether those inferences come from collectives' size, actual functional relations between collectives, or even divisive political rhetoric. This orientation gives us greater purchase on what levers to pull to try to attenuate inter-coalitional conflict.

4.1 Manipulating experienced or implied competition and coordination

If ideologies such as SDO, which imbue people with a chronic tendency to see the world as a competitive place, increase inter-coalitional antagonism, one possibility is that reducing individuals' chronic experience of threat might mitigate said antagonism. In line with this proposition, we found that priming people with secure attachment schemas significantly decreased negative out-group emotions and aggressive inter-coalition behavior (Saleem, Prot, Cikara, Anderson, & Lam, 2013). Across a series of studies employing variants of the guided imagination task (Mikulincer & Shaver, 2001), we randomly assigned participants to write either about an experience with close others or a trip to the grocery story. American participants who recalled a time when someone close to them was available, supportive, and loving reported feeling significantly less intense negative emotions (e.g., fear, anger, disgust) toward Arab people relative to participants in the control condition. This effect was not accounted for by increases in positive mood more generally and replicated when the group was identified as Muslims or Ohio State University students (for the University of Michigan sample). More important, participants primed with a secure attachment schema were less likely to support military and aggressive measures against ISIS members compared with those in the positive mood induction and control conditions. One of the benefits of an approach like this is that it does not harbor the same potential for backfiring, that, say recategorization might. Telling people that they suddenly belong to the same group as "the other side" can threaten group distinctiveness (e.g., Spears, Doosje, & Ellemers, 1997) and dampen motivation for progressive social change of the groups are of differing status (e.g., Saguy, Tausch, Dovidio, & Pratto, 2009). Because it is effectively a *self*-focused strategy, priming attachment security sidesteps these risks.

A more intuitive manipulation is simply to provide people with information that reduces perceptions of other categories as a threat. In a recent suite of experiments, we attempted to leverage positive, achievement-oriented narratives, which emphasize broader contributions to society (and therefore coordination with all collectives) to reduce prejudice and discrimination, specifically toward immigrants (Martinez, Feldman, Feldman, & Cikara, in press). Why immigrants? The U.S. and other western countries have seen massive backlash in response to a perceived influx of immigrants, particularly those who are non-white in recent years. One

driving force of this backlash is the rhetoric—specifically the "criminal narrative"—surrounding the character of immigrants and their impact on residents' lives. Though claims of relatively greater criminality among some immigrant groups are statistically unfounded (Lee & Martinez, 2009), these claims may nevertheless alter the structure of the cognitive representation of "immigrants" as a whole—for example, cleaving immigrants into "good" vs. "bad" subgroups, or more specifically, "white" and "non-white" subgroups (Flores & Schachter, 2018). Of course at the time of data collection these criminal characterizations were already widespread. If participants' representations were subgrouped prior to beginning the experiment, then another way to test whether narratives could alter them was to examine the opposing effect of coordination-emphasizing narratives, e.g., the achievement narrative (e.g., Moffit, Nardon, & Zhang, 2019). From a representation-structure perspective, achievement narratives should counteract subgrouping effects by making all immigrants from all countries more similar to one another, because positive stimuli tend to be rated and represented more similarly than negative stimuli (Alves et al., 2017). Why would subgrouping dampen (or homogenization increase) support for immigrants and immigration? Greater accessibility of negative relative to positive exemplars (Rozin & Royzman, 2001) makes it more likely that people will substitute "bad" immigrants for "immigrants" in general when considering their policy preferences. If all immigrants are the same (and cooperative), negative exemplars are less likely to come to mind and inform policy choices.

Across two experiments we manipulated participants' exposure to criminal, achievement, or struggle-oriented descriptions of immigrants in order to assess how they impact participants' latent representations of four politically salient immigrant groups—Germans, Russians, Syrians, and Mexicans. We included struggle narratives as a control condition to isolate the effect of criminalized narratives above and beyond being negatively-valenced. We then applied a novel analytic technique borrowed from cognitive neuroscience—representational similarity analysis—to extract participants' latent cognitive representations of these immigrant groups and their members.

As predicted, we found that criminal, achievement, and struggle-oriented narratives about different immigrant groups shaped the way people represent these groups and their members; these representations, in turn, informed immigration policy preferences. Most troubling, was that criminal narratives fostered racialized immigrant representations (i.e., creating two clusters of white versus non-white immigrants in trait space)—even among our most egalitarian respondents. Achievement narratives, by contrast, made immigrants from different backgrounds more similar to one another and increased respondents' support for immigration.

Of course, in many cases we cannot simply "prime" away competitiveness or force coordination when there are, or have been, real zero-sum resources at stake (nor would it be desirable to do away entirely with conflict; Cikara & Paluck, 2014). For example, in the Experiments 3a and 3b examining the effect of competition among novel groups on (counter-) empathy we found that the intergroup empathy bias persisted even after competitor threat was decreased (participants learned their team had pulled ahead) or eliminated (participants learned their team had won the problem-solving challenge; Cikara et al., 2014). The recent history of competition without subsequent cues to increased capacity or desire to coordinate carried the empathy gap forward. So what are the alternatives to manipulating the experience or perception of other-coalition competition and status?

4.2 Reducing perceptions of coalitional cohesion via individuation

From a coalitional perspective, the capacity and desire to coordinate (or not) are doing the work of imbuing collectives with the quality of 'groupiness.' Similar to the way that large external magnetic fields align hydrogen nuclei, competition gives rise to the inference that other-coalition members are united in their purpose against 'us.' Thus directly targeting this quality by dismantling perceptions of coalitional *cohesion* carries the potential to change attitudes, emotions, and behavior. For example, in Experiment 4 of the novel groups/empathy paper (Cikara et al., 2014) we found that we could significantly attenuate the intergroup empathy bias by providing participants with visual cues to reduced in-group and out-group distinctiveness (see Figure 9).

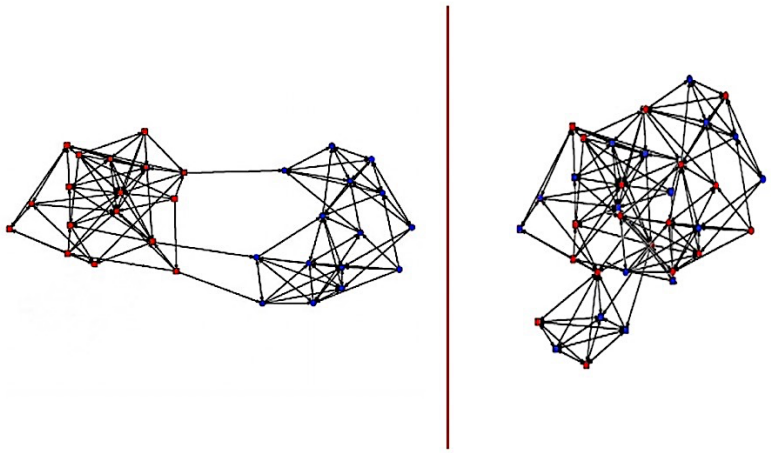

Figure 9. The ostensible social networks of current Rattlers (in red) and Eagles (in blue) players in the high-distinctiveness condition (left), and the low-distinctiveness conditions (right). Adapted from Cikara et al. (2014).

With this goal of dismantling perceptions of coalitional cohesion in mind, we conducted a series of experiments in which we tested whether embedding the good and bad fortune information about same-team versus competitor-team targets in a larger narrative individuated the targets and thereby attenuated the intergroup empathy bias (Bruneau, Cikara, & Saxe, 2015). Indeed it did. Moreover, we found that narratives, which endowed each target with a mind through descriptions of their mental states, were more effective at decreasing empathy bias than narratives focused on their physical descriptions (Figure 10). Critically, poorer memory for group membership (but not memory for other aspects of the scenarios) mediated the relationship between the narrative manipulation and the empathy bias, suggesting that the narratives had their effect by shifting focus away from each target's group membership toward individuating information.

In another line of research, we found that asking people to vividly simulate a helping scenario increased how much people helped opposing political party members (Gaesser, Shimura, & Cikara, 2019). Particularly surprising was that *scene* vividness (rather than the vividness of the person participants imagined helping) and perspective-taking independently drove increases in helping. This suggests that the sensory properties of simulations, especially when they pull attention away from person-specific features like group membership, may play a much larger role in social behavior than we have previously recognized (Vollberg & Cikara, 2018). We have recently replicated this increase in empathy for both same and opposing political party members using an incidental manipulation of episodic simulation that does not direct

people to simulate helping at all, but rather just potentiates one's tendency to engage in episodic simulation (Vollberg, Gaesser, & Cikara, under review).

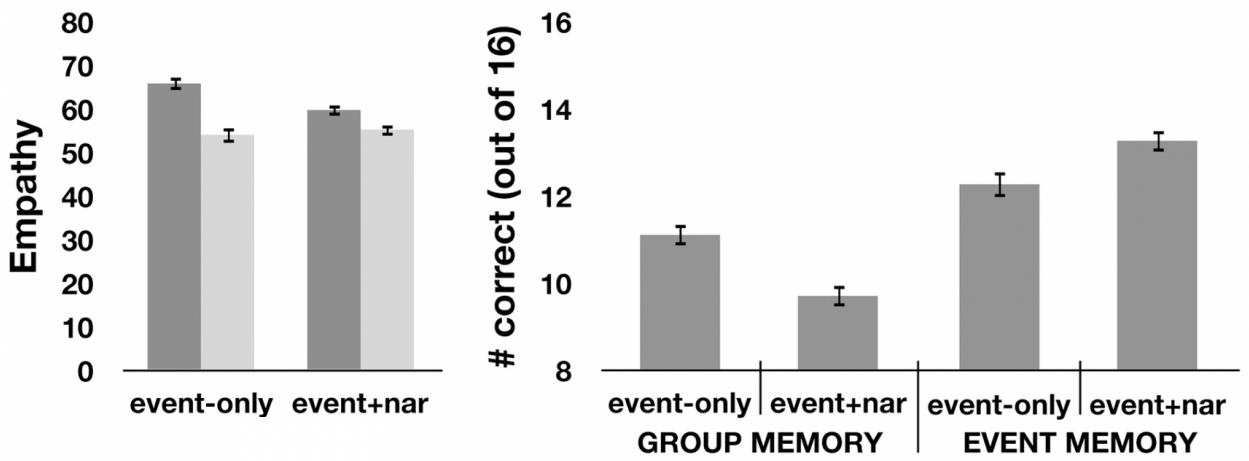

Figure 10. Left: Empathic responses to in-group targets (dark bars) and out-group targets (light bars) in response to good and bad fortunes presented by themselves (event-only) or after a narrative about the target (event + narrative). Right: At the end of the study, participants performed a 2-alternative forced-choice task to recall the group membership of each of the 16 targets (group memory), or the event that happened to each target (event memory). Error bars represent SEM. Adapted from Bruneau et al. (2015).

I would be remiss, however, if I did not emphasize that simply increasing empathy for other-coalition members is not a panacea. One common assumption is that exercises designed to increase empathy should eliminate empathic failures and their assumed consequences (e.g., harm), but the story is not so simple (for reviews, see Zaki & Cikara, 2015; Weisz & Cikara, in press). Our findings strongly challenge that assumption. We find that neglectful behavior and policy preferences are better predicted by the size of the *gap* between in-group and out-group empathy, rather than participants' absolute levels of empathy. For example, across three studies we found that American respondents cared more about fellow Americans suffering misfortunes relative to Arabs; Hungarians were more empathic toward their countrymen than Muslim refugees; and the same with Greeks toward Germans. And across all three studies, people's empathy gaps—controlling for participants' trait empathic concern—predicted their perspectives on policy: how much Americans approved of Arab immigration, how many asylum seekers Hungary should accept, and how much Greece should help if a natural disaster hit Germany (Bruneau, Cikara, & Saxe, 2017). Therefore, interventions or programs aimed at increasing overall empathy (e.g., generalized compassion training) may have little or no effect on increasing inter-coalition harmony or so long as that gap is maintained.

One final intriguing possibility is that we can fight a group-level cognitive bias with an individuating cognitive bias: specifically in this case the good-true-self bias, which falls specifically out of reasoning about the essence of an individual person. Mounting evidence indicates that people exhibit a robust, invariant tendency to believe that inside every individual there is a 'good true self' calling every person to behave in morally virtuous ways and that this essence is separate from a person's superficial features (Strohminger et al., 2017). This bias is present across cultures, perspectives (first versus third), and individual differences, and appears to be rooted in the basic cognitive tendency to assume that *all* entities have deep, unobservable,

inherent properties that comprise their true nature (DeFreitas et al., 2017). If this is the case, then even threatening category members should be judged as having good true selves, deep down. A positive bias that falls out of thinking about the essence of an individual person could be leveraged to reduce a negative bias that falls out of thinking about the nature of a threatening coalition.

Across three experiments we tested whether Americans believed that an American, an Arab, and an Arab in the U.S. all contain good true selves, deep down, to equal extents (DeFreitas & Cikara, 2017). Not only did our participants attribute good true selves across these targets in equal measure, we found that asking Americans to reflect on individual Arab targets' true selves first made them less prejudiced toward and less threatened by Arab people in general and more likely to donate money to the Syrian Arab Red Crescent charity (relative to people who completed prejudice/donation judgments prior to making true self judgments). We submit that thinking about whether an agent's behavior reflected their true versus surface-self led to a particularly strong form of individuation, which in turn led to more nuanced representations of all people as possessing multiple layers (i.e., surface self could go either way but the true self is good) rather than merely characterizing people as "us" (good) versus "them" (bad).

5. CONCLUSION

To summarize:

- The contemporary intergroup literature has emphasized the role of category-membership over coalitional structure; however research treating demographic categories as purposive groups will often run into explanatory limitations *and* may ironically end up reinforcing stereotypes.
- We should focus much more instead on the conditions under which coalitional psychology gets activated: the recognition that another is able and willing (or not) to engage in behavior having accounted for your and others' welfare.
    - Even in the absence of threat or competition, the inference of coordination difficulty or improbability may be sufficient to mark someone as an out-group member. Coordination ease (or difficulty) may or may not track with shared category membership.
    - People of all ages are sensitive to how well agents coordinate not just with themselves but with others in the environment, indicating that people are prone to building representations of coordinated coalitions—or social structures—out in the world rather than just egocentric, dyadic similarities or interdependencies.
    - Neural responses associated with generalized "us" vs. "them" representations are consistent with the hypothesis that salience, specifically functional significance or evaluation (i.e., will this stimulus help me or not?), is the primary dimension of distinction.
    - As predicted by a coalitional account, stereotypes and prejudices change because people are sensitive to generalized group features that signal threat or coordination, invariant to the groups in question (e.g., members of the *largest* minoritized group in a county were significantly more likely to be targeted with hate crimes relative to when their own category ranked second or lower in the minority group size distribution in the same county).

- Once coalitional psychology has been activated there is a whole cascade of consequences for our social preferences, emotions, and attributions. However, the nature of these consequences will hinge on the particulars of the functional relations between coalitions.
    - People are more willing to sacrifice groups that are stereotyped as unable and unwilling to coordinate in the absence of any group labels.
    - Acting as part of a competitive collective can reduce the salience of one's own moral standards and, in turn, enable out-group harm.
    - Inter-coalitional competition not only reduces empathy but also increases pleasure in response to the other side's suffering, i.e., Schadenfreude: in sports contexts, when targets are merely stereotyped as competitive and able, and when equally able novel groups are placed in competition with one another. This emotional profile is associated with greater willingness to harm the other side.
    - Preliminary findings suggest that emotions like Schadenfreude may serve to motivate preferences for out-group harm via (i) reinforcement learning-like updating mechanisms and (ii) observers' ideologies: specifically Social Dominance Orientation.
    - Cues to coalition competition and status also shape our attributions and expectations of emotion to other-coalition members. People judge same-team faces as more positive than competitor-faces, regardless of emotion, across multiple modes of measurement. More important for conflict escalation: people significantly overestimate how outraged competitive coalitions will be toward their own group's behaviors, which is associated with reduced estimates of of the possibility of coordination.
- One of the primary benefits of a coalitional account is that it highlights the specific cues that drive the cleaving of collectives into "in-groups" and "out-groups."
    - Reducing individuals' chronic experience of threat via secure attachment primes or providing people with information that reduces perceptions of other categories as threatening mitigates inter-coalitional antagonism.
    - Reducing perceptions of coalitional cohesion—via individuation, episodic simulation, or the good-true-self bias—also increases empathy and helping behavior toward competitive, threatening out-groups.

My primary goal here was to catalog some of the evidence we have to indicate that much of what we understand as intergroup conflict stems not from category membership or features we believe are intrinsic to said categories but rather from our recognition of one another's capacity for and likelihood of coordination: not only with oneself but with others. The principal strength of the coalitional approach is that it allows us to make predictions about novel intergroup contexts and about how intergroup dynamics may change over time rather than having to appeal post-hoc to intergroup-specific factors to explain conflict.

That said, I want to be clear that my aim is not to dismiss the roles of specific group histories or group-bound stereotypes. Nor do I want to claim histories and stereotypes have no explanatory power. Rather my aim is to ask: where do those histories and stereotypes come from? What would a unifying framework of "group" psychology look like? To answer these questions we have to abstract away from many of the details that are category-bound. I also think it is important to remember that so many of the categories that intergroup conflict researchers study are those that are imposed by us on the very populations we are interested in understanding (whether or not their constituent members consent to those category assignments). These priors

are at best placeholders and at worst misguided structures preventing us from gaining a deeper understanding of the causal factors and constructs at work in conflict.

There are many exciting next steps to take in this line of thinking, many of which I have already highlighted. Very broadly, our field must grapple with the notion that we lack a scientific definition of the generalized concept of "group." Instead we appeal to attributes that tend to characterize collectives we perceive as "groups" to then define groups as construct (Pietraszewski, in press). How would we program a robot to recognize a group? To figure out to which group it belongs? We have already seen that similarity is insufficient. So what are the correct inputs? My hope is that the framework and findings reviewed here will help to move this line of questioning forward.

References

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison Wesley.

Amodio, D. M., & Cikara, M. (in press). Social neuroscience of prejudice. *Annual Review of Psychology.*

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Balliet, D., Tybur, J. M., & Van Lange, P. A. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review*, *21*(4), 361-388.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological bulletin*, *140*(6), 1556.

Bandura, A., 1999. Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review, 3, 193–209.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*, *76*, 412-427.

Batson, C.D. (2009). These things called empathy: Eight related but distinct phenomena. In J. Decety & W.J. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–15). Cambridge, MA: MIT Press.

Batson, C.D., & Ahmad, N.Y. (2009). Using empathy to improve intergroup attitudes and relations. *Social Issues and Policy Review, 3*, 141.

Bauer, M., Cahlíková, J., Chytilová, J., & Želinský, T. (2018). Social contagion of ethnic hostility. *Proceedings of the National Academy of Sciences*, *115*(19), 4881-4886.

Beer, J.S., Knight, R.T., D'Esposito, M. (2006). Controlling the integration of emotion and cognition: the role of frontal cortex in distinguishing helpful from hurtful emotional information. *Psychological Science, 17*, 448–53.

Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: eliminate the negative, accentuate the positive. *Journal of personality and social psychology*, *102*(6), 1214.

Bhui, R., & Gershman, S. J. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological review*, *125*(6), 985.

Bishop, S., Duncan, J., Brett, M., Lawrence, A.D. (2004). Prefrontal cortical function and anxiety: controlling attention to threat-related stimuli. *Nature Neuroscience, 7,* 184–8.

Blalock, H. M. (1957). Per cent non-white and discrimination in the South. *American Sociological Review*, *22*(6), 677-682.

Bobo, L. (1983). Whites' opposition to busing: Symbolic racism or realistic group conflict?. *Journal of personality and social psychology*, *45*(6), 1196.

Boecker, L., Likowski, K. U., Pauli, P., & Weyers, P. (2015). The face of Schadenfreude: Differentiation of joy and Schadenfreude by electromyography. *Cognition and Emotion*, *29*(6), 1117-1125.

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and social psychology review*, *7*(2), 129-145.

Brandt, M. J., & Crawford, J. T. (2020). Worldview conflict and prejudice. In *Advances in Experimental Social Psychology* (Vol. 61, pp. 1-66). Academic Press.

Braun, R. (2014). Religious minorities and resistance to genocide: the collective rescue of Jews in the Netherlands during the Holocaust. In *APSA 2014 Annual Meeting Paper*.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological bulletin*, *86*(2), 307.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, *17*(5), 475-482.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and out-group hate? Journal of Social Issues, 55(3), 429–444.

Brewer, M. B. (2000). Superodinate goals versus superordinate identity as bases of intergroup cooperation.

Brewer, M. B. (2001). In-group identification and intergroup conflict: When does in-group love become out-group hate? In R. D. Ashmore, L. Jussim, & D. Wilder (Eds.), Social identity, intergroup conflict, and conflict reduction (pp. 17–41). New York: Oxford University Press.

Brick, C., Hood, B., Ekroll, V., & de-Wit, L. (in press). Illusory essences: A bias holding back theorizing in psychological science. *Perspectives on Psychological Science*.

Brown, S. L., & Schwartz, G. E. (1980). Relationships between facial electromyography and subjective experience during affective imagery. *Biological psychology*, *11*(1), 49-62.

Bruneau, E. G., Cikara, M., & Saxe, R. (2015). Minding the gap: Narrative descriptions about mental states attenuate parochial empathy. PLoS ONE, 10(10): e0140838.

Bruneau, E. G., Cikara, M., & Saxe, R. (2017). Parochial empathy predicts the reduced altruism and the endorsement of passive harm. *Social Psychological and Personality Science*.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral science*, *3*(1), 14.

Campbell, D. T. (1965). Ethnocentric and other altruistic motives. In *Nebraska symposium on motivation* (Vol. 13, pp. 283-311).

Chang, L. W., Gershman, S. J., Cikara, M. (2019). Comparing value-coding models of context-dependence in social choice. *Journal of Experimental Social Psychology, 85*, 103847.

Chang, L. W., & Cikara, M. (2018). Social decoys: Leveraging choice architecture to alter social preferences. *Journal of Personality and Social Psychology, 115*, 206-223.

Chang, L. W., Krosch, A. R., & Cikara, M. (2016). Effects of intergroup threat on mind, brain, and behavior. *Current Opinion in Psychology, 11*, 69-73.

Chester, D. S. (2017). The role of positive affect in aggression. *Current Directions in Psychological Science*, *26*(4), 366-370.

Chester, D. S., & DeWall, C. N. (2016). The pleasure of revenge: retaliatory aggression arises from a neural imbalance toward reward. *Social cognitive and affective neuroscience*, *11*(7), 1173-1182.

Cikara, M. (2015). Intergroup Schadenfreude: Motivating participation in collective violence. *Current Opinion in Behavioral Sciences, 3*, 12-17.

Cikara, M., Botvinick, M. M., & Fiske, S. T. (2011). Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological Science, 22*, 306-313.

Cikara, M., Bruneau, E. G., & Saxe, R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science, 20*, 149-153.

Cikara, M., Bruneau, E. G., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology, 55*, 110-125.

Cikara, M., Farnsworth, R. A., Harris, L. T., & Fiske, S. T. (2010). On the wrong side of the trolley track: Neural correlates of relative social valuation. *Social Cognitive and Affective Neuroscience, 5*, 404-413.

Cikara, M., & Fiske, S. T. (2011). Bounded empathy: Neural responses to outgroup targets' (mis)fortunes. *Journal of Cognitive Neuroscience, 23*, 3791-3803.

Cikara, M., & Fiske, S. T. (2012). Stereotypes and Schadenfreude: Affective and physiological markers of pleasure at others' misfortunes. *Social Psychological and Personality Science, 3*, 63-71.

Cikara, M., & Fiske, S. T. (2013). Their pain, our pleasure: Stereotype content and Schadenfreude. *Annals of the New York Academy of Sciences, 1299*, 52-59.

Cikara, M., & Fiske, S. T. (2014). Stereotypes and Schadenfreude. In W. van Dijk & J. Ouwerkerk (Eds.) *Schadenfreude* (pp. 151-169). Cambridge: Cambridge University Press.

Cikara, M., Fouka, V., & Tabellini, M. (under review). Hate crime increases with minortized group rank.

Cikara, M., Jenkins, A., Dufour, N., & Saxe, R. (2014). Reduced self-referential neural response during intergroup competition predicts later willingness to harm a competitor. *NeuroImage, 96*, 36-43.

Cikara, M., & Paluck, E. L. (2013). When going along gets you nowhere and the upside of conflict behaviors. *Social and Personality Psychology Compass, 7*, 559-571.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science, 9*, 245-274.

Cikara, M., Van Bavel, J. J., Ingbretsen, Z., & Lau, T. (2017). Decoding "us" and "them:" Neural representations of generalized group concepts. J*ournal of Experimental Psychology: General, 146*, 621-631.

Cohen, T. R., & Insko, C. A. (2008). War and peace: Possible approaches to reducing intergroup conflict. *Perspectives on Psychological Science*, *3*(2), 87-93.

Cohen, T. R., Montoya, R. M., & Insko, C. A. (2006). Group morality and intergroup relations: Cross-cultural and experimental evidence. *Personality and Social Psychology Bulletin*, *32*(11), 1559-1572.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*(1), 37.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. In *Advances in experimental social psychology* (Vol. 56, pp. 131-199). Academic Press.

Cortland, C. I., Craig, M. A., Shapiro, J. R., Richeson, J. A., Neel, R., & Goldstein, N. J. (2017). Solidarity through shared disadvantage: Highlighting shared experiences of discrimination improves relations between stigmatized groups. *Journal of Personality and Social Psychology*, *113*(4), 547.

Craig, M. A., & Lee, M. M. (under review). Status-based coalitions: Hispanic growth affects Whites' perceptions of political support from Asian Americans.

Craig, M. A., Rucker, J. M., & Richeson, J. A. (2018). The pitfalls and promise of increasing racial diversity: Threat, contact, and race relations in the 21st century. *Current Directions in Psychological Science, 27*(3), 188-193.

Crandall, C. S., Eshleman, A., & O'brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of personality and social psychology*, *82*(3), 359.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, *92*(4), 631.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology, 8*, 377–383.

de Bruijn, E. R., de Lange, F. P., von Cramon, D. Y., & Ullsperger, M. (2009). When errors are rewarding. *Journal of Neuroscience*, *29*(39), 12183-12186.

Decety, J. (2011). Dissecting the neural mechanisms mediating empathy. *Emotion review*, *3*(1), 92-108.

De Dreu, C. K., Gross, J., Fariña, A., & Ma, Y. (2020). Group Cooperation, Carrying-Capacity Stress, and Intergroup Conflict. *Trends in Cognitive Sciences*.

De Freitas, J., Cikara, M., Grossman, I., & Schlegel, R. (2017). Origins of the belief in good true selves. Trends in Cognitive Sciences, 21, 634-636.

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology, 74*, 307-316.

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of cognitive Neuroscience*, *24*(8), 1742-1752.

Derks, B., Ellemers, N., Van Laar, C., & de Groot, K. (2011). Do sexist organizational cultures create the queen bee? *British Journal of Social Psychology, 50*, 519–535

Derks, B., Van Laar, C., Ellemers, N., & de Groot, K. (2011). Gender-bias primes elicit queen-bee responses among senior policewomen. *Psychological Science, 22*, 1243–1249.

Derks, B., Van Laar, C., & Ellemers, N. (2016). The queen bee phenomenon: Why women leaders distance themselves from junior women. *The Leadership Quarterly*, *27*(3), 456-469.

Ellemers, N., De Gilder, D., & Haslam, S. A. (2004). Motivating individuals and groups at work: A social identity perspective on leadership and group performance. *Academy of Management review*, *29*(3), 459-478.

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and social identity. *Annual review of psychology*, *53*(1), 161-186.

Esses, V. M., Jackson, L. M., & Armstrong, T. L. (1998). Intergroup competition and attitudes toward immigrants and immigration: An instrumental model of group conflict. *Journal of social issues*, *54*(4), 699-724.

Esses, V. M., Jackson, L. M., Dovidio, J. F., & Hodson, G. (2005). Instrumental relations among groups: Group competition, conflict, and prejudice. *On the nature of prejudice: Fifty years after Allport*, 227-243.

FeldmanHall, O., Mobbs, D., Hiscox, L., Navrady, L., Dalgleish, T. (2012). What We Say and What We Do: The Relationship Between Real and Hypothetical Moral Choices. *Cognition, 123,* 434-441.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67-73.

Fiske, S. T. (2018b). *Social beings: Core motives in social psychology*. John Wiley & Sons.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, *82*(6), 878.

Flores, R. D., & Schachter, A. (2018). Who are the "Illegals"? The Social Construction of Illegality in the United States. *American Sociological Review*, *83*(5), 839–868.

Fossett, M. A., & Kiecolt, K. J. (1989). The relative size of minority populations and white racial attitudes. *Social Science Quarterly*, *70*(4), 820.

Fouka, V., Mazumder, S., & Tabellini, M. (2019). *From Immigrants to Americans: Race and Assimilation during the Great Migration* (No. 19-018). Harvard Business School.

Gaertner, S. L., Dovidio, J. F., Banker, B. S., Houlette, M., Johnson, K. M., & McGlynn, E. A. (2000). Reducing intergroup conflict: From superordinate goals to decategorization, recategorization, and mutual differentiation. *Group Dynamics: Theory, Research, and Practice*, *4*(1), 98.

Gaertner, S. L., Mann, J., Murrell, A., & Dovidio, J. F. (1989). Reducing intergroup bias: The benefits of recategorization. *Journal of personality and social psychology*, *57*(2), 239.

Gaertner, L., & Schopler, J. (1998). Perceived ingroup entitativity and intergroup bias: An interconnection of self and others. *European Journal of Social Psychology*, *28*(6), 963-980.

Gaesser, B., Shimura, Y., & Cikara, M. (2019). Episodic simulation reduces intergroup bias in prosocial intentions and behavior. Journal of Personality and Social Psychology.

Glaser, J. M. (1994). Back to the black belt: Racial environment and white racial attitudes in the South. *The Journal of Politics*, *56*(1), 21-41.

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. Current Directions in Psychological Science, 29, 460-466.

Halevy, N., Bornstein, G., & Sagiv, L. (2008). "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological science*, *19*(4), 405-411.

Halevy, N., Chou, E. Y., Cohen, T. R., & Bornstein, G. (2010). Relative deprivation and intergroup competition. *Group Processes & Intergroup Relations*, *13*(6), 685-700.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological review*, *103*(2), 336.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me= bad: Infants prefer those who harm dissimilar others. *Psychological science*, *24*(4), 589-594.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the national academy of sciences*, *108*(50), 19931-19936.

Han, S. (2018). Neurocognitive basis of racial ingroup bias in empathy. *Trends in Cognitive Sciences*, *22*(5), 400-421.

Harris, L. T., Cikara, M., & Fiske, S. T. (2008). Envy as predicted by the stereotype content model: A volatile ambivalence. In R. Smith (Ed.), Envy: Theory and research (pp. 131-147). New York: Oxford University Press.

Hauser, M., Cushman, F., Young, L., Jin, R.K., Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language, 22*, 1–21.

Heider, F. (1958). The psychology of interpersonal relations. New York: Wiley.

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron*, *68*(1), 149-160.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual review of psychology*, *53*(1), 575-604.

Hjerm, M. (2009). Anti-immigrant attitudes and cross-municipal variation in the proportion of immigrants. *Acta sociologica*, *52*(1), 47-62.

Hood III, M. V., & Morris, I. L. (1997). ¿ Amigo o enemigo?: Context, attitudes, and Anglo public opinion toward immigration. *Social Science Quarterly*, 309-323.

Hoogland, C. E., Schurtz, D. R., Cooper, C. M., Combs, D. J., Brown, E. G., & Smith, R. H. (2015). The joy of pain and the pain of joy: In-group identification predicts Schadenfreude and gluckschmerz following rival groups' fortunes. *Motivation and Emotion*, *39*(2), 260-281.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, *9*(1), 90-98.

Huddy, L. (2001). From social to political identity: A critical examination of social identity theory. *Political psychology*, *22*(1), 127-156.

Hudson, S. T. J., Cikara, M., & Sidanius, J. (2019). Preference for hierarchy is associated with reduced empathy and increased counter-empathy, especially toward out-groups. Journal of Experimental Social Psychology, 85, 103871.

Hudson, S. T. J., Cikara, M., & Sidanius, J. (under review). Preference for hierarchy is related to the motivation to feel less empathy and more Schadenfreude toward low status people.

Humes, K. R., Jones, N. A., & Ramirez, R. (2011). Overview of Race and Hispanic Origin: 2010. The Census Bureau, Washington, DC.

Ichikawa, K., Moore, W., & Cikara, M. (2020). Neural Correlates of Latent Preferences for out-Group Harm. *APS Virtual Poster Showcase*.

Insko, C. A., Wildschut, T., & Cohen, T. R. (2013). Interindividual–intergroup discontinuity in the prisoner's dilemma game: How common fate, proximity, and similarity affect intergroup competition. *Organizational Behavior and Human Decision Processes*, *120*(2), 168-180.

Ito, T. A., & Bartholow, B. D. (2009). The neural correlates of race. *Trends in cognitive sciences*, *13*(12), 524-531.

Jenkins, A. C., & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social neuroscience*, *6*(3), 211-218.

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence*. John Wiley & Sons.

Keltner, D. (2009). Born to be good: The science of a meaningful life. New York: Norton.

Kinder, D. R., & Sears, D. O. (1981). Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of personality and social psychology*, *40*(3), 414.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577-12580.

Kinzler, K. D., Shutts, K., DeJesus, J., & Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social cognition*, *27*(4), 623-634.

Koch, A., Imhoff, R., Unkelbach, C., Nicolas, G., Fiske, S., Terache, J., ... & Yzerbyt, V. (2020). Groups' warmth is a personal matter: Understanding consensus on stereotype dimensions reconciles adversarial models of social evaluation. *Journal of Experimental Social Psychology, 89*, 103995.

Krosch, A. R., Tyler, T. R., & Amodio, D. M. (2017). Race and recession: Effects of economic scarcity on racial discrimination. *Journal of personality and social psychology*, *113*(6), 892.

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature neuroscience*, *15*(7), 940-948.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, *14*(5), 402-408.

Lazerus, T., Ingbretsen, Z., Stolier, R., Freeman, J. B., & Cikara, M. (2016). Positivity bias in judgments of in-group members' emotions. Emotion, 16, 1117-1125.

Lau, T., Gershman, S. J., & Cikara, M. (2020). Social structure learning in human anterior insula. eLife, 9, e53162. (See "Spotlight" feature on this paper in Trends in Cognitive Sciences)

Lau, T., Morewedge, C. K., & Cikara, M. (2016). Overcorrection for social categorization information moderates impact bias in affective forecasting. Psychological Science, 27, 1340-1351.

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. Journal of Experimental Psychology: General, 147, 1881-1891.

Leach, C. W., & Spears, R. (2008). " A vengefulness of the impotent": The pain of in-group inferiority and Schadenfreude toward successful out-groups. *Journal of personality and social psychology*, *95*(6), 1383.

Leach, C. W., & Spears, R. (2009). Dejection at in-group defeat and Schadenfreude toward second-and third-party out-groups. *Emotion*, *9*(5), 659.

Leach, C. W., Spears, R., Branscombe, N. R., & Doosje, B. (2003). Malicious pleasure: Schadenfreude at the suffering of another group. *Journal of personality and social psychology*, *84*(5), 932.

Lee, M. T., & Martinez, R. (2009). Immigration reduces crime: an emerging scholarly consensus. *Sociology of Crime, Law and Deviance*, *13*, 3–16.

Legrand, R. (2013). Successful aggression as the reinforcer for runway behavior of mice. Psychonomic Science, 20(5), 303–305.

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. Nature Human Behaviour, 4, 279–286.

Lees, J., & Cikara, M. (in press). Understanding and combating misperceived polarization. Philosophical Transactions of the Royal Society B: Biological Sciences.

Lenton, A.P. (2002). The price of prejudice: social categories influence monetary value of life. Dissertation Abstracts International: Section B: The Sciences and Engineering, 63, 1–139.

LeVine, R. A., & Campbell, D. T. (1972). Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior.

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. Journal of Personality and Social Psychology, 78, 223–246.

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

Mackie, D. M., & Smith, E. R. (2015). Intergroup emotions. In *APA handbook of personality and social psychology, Volume 2: Group processes.* (pp. 263-293). American Psychological Association.

Martinez, J. E., Feldman, L., Feldman, M., & Cikara, M. (in press). Narratives shape cognitive representations of immigrant groups and policy preferences. Psychological Science.

Martinez, J. E., & Paluck, E. L. (2020). Quantifying shared and idiosyncratic judgments of racism in social discourse. PsyArXiv.

McDoom, O. S. (2012). The psychology of threat in intergroup conflict: emotions, rationality, and opportunity in the Rwandan genocide. *International Security*, *37*(2), 119-155.

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

Meier, B. P., & Hinsz, V. B. (2004). A comparison of human aggression committed by groups and individuals: An interindividual–intergroup discontinuity. *Journal of Experimental Social Psychology*, *40*(4), 551-559.

Mikulincer, M., & Shaver, P. R. (2001). Attachment theory and intergroup bias: Evidence that priming the secure base schema attenuates negative reactions to out-groups. *Journal of personality and social psychology*, *81*(1), 97.

Moffitt, U. E., Nardon, L., & Zhang, H. (2019). Becoming Canadian: Immigrant narratives of professional attainment. *International Journal of Intercultural Relations*.

Molenberghs, P., & Morrison, S. (2014). The role of the medial prefrontal cortex in social categorization. *Social cognitive and affective neuroscience*, *9*(3), 292-296.

Morrison, S., Decety, J., & Molenberghs, P. (2012). The neuroscience of group membership. *Neuropsychologia*, *50*(8), 2114-2120.

Moya, C., & Scelza, B. (2015). The effect of recent ethnogenesis and migration histories on perceptions of ethnic group stability. *Journal of Cognition and Culture*, *15*(1-2), 131-173.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*(2), 103-122.

Neel, R., & Shapiro, J. R. (2012). Is racial bias malleable? Whites' lay theories of racial bias predict divergent strategies for interracial interactions. *Journal of Personality and Social Psychology*, *103*(1), 101.

Neuberg, S. L., & Cottrell, C. A. (2006). Evolutionary bases of prejudices. *Evolution and social psychology*, 163-187.

O'Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology, 14*, 769-776.

Ofosu, E. K., Chambers, M. K., Chen, J. M., & Hehman, E. (2019). Same-sex marriage legalization associated with reduced implicit and explicit antigay bias. *Proceedings of the National Academy of Sciences*, *116*(18), 8846-8851.

Oliver, E. J. & Mendelberg, T. (2000). Reconsidering the Environmental Determinants of White Racial Attitudes. *American Journal of Political Science, 44*, 574–89.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233-248.

Perdue, C. W., Dovidio, J. F., Gurtman, M. B., & Tyler, R. B. (1990). Us and them: social categorization and the process of intergroup bias. *Journal of personality and social psychology*, *59*(3), 475.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of personality and social psychology*, *64*(3), 467.

Pettibone, J. C., & Wedell, D. H. (2000). Examining models of nondominated decoy effects across judgment and choice. *Organizational behavior and human decision processes*, *81*(2), 300-328.

Pettigrew, T. F. (1957). Demographic correlates of border-state desegregation. *American Sociological Review*, *22*(6), 683-689.

Pietraszewski, D. (2013). What is group psychology? Adaptations for mapping shared intentional stances. In M. Banaji, & S. Gelman (Eds). *Navigating the social world: What infants, children, and other species can teach us*, p. 253-257. New York: Oxford.

Pietraszewski, D. (2016). How the mind sees coalitional and group conflict: The evolutionary invariances of coalitional conflict dynamics. *Evolution and Human Behavior, 37*, 470-480.

Pietraszewski, D. (2020). Intergroup processes: Principles from an evolutionary perspective. In *Social Psychology: Handbook of Basic Principles.* In P. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds). (pp. 373-391). New York: Guilford.

Pietraszewski, D. (in press). Toward a computational theory of social groups: A finite set of cognitive primitives for representing any and all social groups in the context of conflict. *Brain and Behavioral Sciences*.

Pietraszewski, D. (in press, b). The correct way to test the hypothesis that racial categorization is a byproduct of an evolved alliance-tracking capacity. *Scientific Reports*

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PLoS ONE 9(2): e88534.*

Pietraszewski, D., & Schwartz, A. (2014). Evidence that accent is a dedicated dimension of social categorization, not a byproduct of coalitional categorization. *Evolution and Human Behavior, 35*, 51-57.

Quillian, L. (1996). Group threat and regional change in attitudes toward African-Americans. *American Journal of Sociology*, *102*(3), 816-860.

Rabbie, J. M., Schot, J. C., & Visser, L. (1989). Social identity theory: A conceptual and empirical critique from the perspective of a behavioural interaction model. European Journal of Social Psychology, 19, 171–202.

Rand, D. G., Pfeiffer, T., Dreber, A., Sheketoff, R. W., Wernerfelt, N. C., & Benkler, Y. (2009). Dynamic remodeling of in-group bias during the 2008 presidential election. Proceedings of the National Academy of Sciences of the United States of America, 106, 6187–6191.

Rhodes, M., & Chalik, L. (2013). Social categories as markers of intrinsic interpersonal obligations. *Psychological Science, 24*, 999–1006.

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., … Cikara, M., Folke, T. (under revision). The general fault in our fault lines. *Nature Human Behaviour*.

Saguy, T., Tausch, N., Dovidio, J. F., & Pratto, F. (2009). The irony of harmony: Intergroup contact can produce false expectations for equality. *Psychological Science*, *20*(1), 114-121.

Saleem, M., Prot, S., Cikara, M., Lam, C.P., Anderson, C.A., & Jelic, M. (2015). Cutting Gordian Knots: Reducing prejudice through attachment security. Personality and Social Psychology Bulletin 41, 1560-1574.

Schaller, M., & Neuberg, S. L. (2012). Danger, disease, and the nature of prejudice (s). In *Advances in experimental social psychology* (Vol. 46, pp. 1-54). Academic Press.

Scheepers, D., Spears, R., Doosje, B., & Manstead, A. S. (2006). Diversity in in-group bias: Structural factors, situational features, and social functions. *Journal of personality and social psychology*, *90*(6), 944.

Schlueter, E., & Scheepers, P. (2010). The relationship between outgroup size and anti-outgroup attitudes: A theoretical synthesis and empirical test of group threat-and intergroup contact theory. *Social Science Research*, *39*(2), 285-295.

Sherif, M. (1948). The necessity of considering current issues as part and parcel of persistent major problems: Illustrated by the problem of prejudice. *International Journal of Opinion and Attitude Research, 2*, 63–68.

Sherif, M. (1966). In common predicament: Social psychology of intergroup conflict and cooperation. Boston, MA: Houghton Mifflin.

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R. & Sherif, C. W. (1961). Intergroup cooperation and competition: The Robbers Cave experiment. University Book Exchange.

Sherif, M., & Sherif, C. W. (1953). Groups in harmony and tension; an integration of studies of intergroup relations.

Sidanius, J., & Pratto, F. (2001). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of consumer research*, *16*(2), 158-174.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466-469.

Smith, R. H. (2000). Assimilative and contrastive emotional reactions to upward and downward social comparisons. In *Handbook of social comparison* (pp. 173-200). Springer, Boston, MA.

Smith, R.H., Powell, C.A.J., Combs, D.J.Y., & Schurtz, D.R. (2009). Exploring the when and why of Schadenfreude. *Social and Personality Psychology Compass, 3*, 530–546.

Smith, R. H., Turner, T. J., Garonzik, R., Leach, C. W., Urch-Druskat, V., & Weston, C. M. (1996). Envy and Schadenfreude. *Personality and Social Psychology Bulletin*, *22*(2), 158-168.

Spears, R., Doosje, B., & Ellemers, N. (1997). Self-stereotyping in the face of threats to group status and distinctiveness: The role of group identification. *Personality and social psychology bulletin*, *23*(5), 538-553.

Staines, G., Tavris, C., & Jayaratne, T. E. (1974). The queen bee syndrome. Psychology Today, 7(8), 55–60

Stephan, W. G., & Stephan, C. W. (2017). Intergroup threat theory. *The international encyclopedia of intercultural communication*, 1-12.

Stürmer, S., Snyder, M., Kropp, A., & Siem, B. (2006). Empathy-motivated helping: The moderating role of group membership. *Personality and Social Psychology Bulletin*, *32*(7), 943-956.

Stürmer, S., Snyder, M., & Omoto, A. M. (2005). Prosocial emotions and helping: the moderating role of group membership. *Journal of personality and social psychology*, *88*(3), 532.

Tajfel, H., (1982). Social Identity and Intergroup Relations. Cambridge University Press, Cambridge, England.

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.

Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worschel (Eds.), The social psychology of intergroup relations (pp. 33–47). Pacific Grove, CA: Brooks/Cole Publishing.

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., & Okubo, Y. (2009). When your gain is my pain and your pain is my gain: neural correlates of envy and Schadenfreude. *Science*, *323*(5916), 937-939.

Tomasello, M. (2009). Why we cooperate. Cambridge, MA: MIT Press.

Tomov, M. S., Dorfman, H. M., & Gershman, S. J. (2018). Neural computations underlying causal structure learning. *Journal of Neuroscience*, *38*(32), 7143-7157.

Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and collective: Cognition and social context. *Personality and social psychology bulletin*, *20*(5), 454-463.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision making*, *4*(6), 479-491.

Vachon, D. D., Lynam, D. R., & Johnson, J. A. (2014). The (non) relation between empathy and aggression: Surprising results from a meta-analysis. *Psychological bulletin*, *140*(3), 751.

Van Bavel, J. J., & Cunningham, W. A. (2010). A social neuroscience approach to self and social categorisation: A new look at an old issue. *European review of social psychology*, *21*(1), 237-284.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: A functional magnetic resonance imaging investigation. *Psychological science*, *19*(11), 1131-1139.

Van Dijk, W. W., Ouwerkerk, J. W., Goslinga, S., & Nieweg, M. (2005). Deservingness and Schadenfreude. *Cognition and Emotion*.

Van Dijk, W. W., Ouwerkerk, J. W., Smith, R. H., & Cikara, M. (2015). The role of self-evaluation and envy in Schadenfreude. European Review of Social Psychology, 26, 247-282.

Van Erp, A. M., & Miczek, K. A. (2000). Aggressive behavior, increased accumbal dopamine, and decreased cortical serotonin in rats. *Journal of Neuroscience*, *20*(24), 9320-9325.

Van Hiel, A., Hautman, L., Cornelis, I., & De Clercq, B. (2007). Football hooliganism: Comparing self-awareness and social identity theory explanations. *Journal of community & applied social psychology*, *17*(3), 169-186.

Vollberg, M., & Cikara, M. (2018). The neuroscience of intergroup emotions. Current Opinions in Psychology, 24, 48-52.

Vollberg, M., Gaesser, B., & Cikara, M. (in press). Activating episodic simulation increases affective empathy. Cognition.

Waters, M. C., & Eschbach, K. (1995). Immigration and Ethnic and Racial Inequality in the United States. *Annual Review of Sociology, 21*, 419-46.

Wedell, D. H. (1994). Contextual contrast in evaluative judgments: A test of pre- versus post integration models of contrast. *Journal of Personality and Social Psychology, 66*, 1007-1019.

Weisz, E., & Cikara, M. (in press). Strategic regulation of empathy. Trends in Cognitive Sciences.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251-1263.

Wong, C., Bowers, J., Williams, T., & Simmons, K. D. (2012). Bringing the Person Back In: Boundaries, Perceptions, and the Measurement of Racial Context. *The Journal of Politics*, *74*, 1153–1170.

Woolf, L. M., & Hulsizer, M. R. (2004). Hate Groups for Dummies: How to Build a Successful Hate-Group. *Humanity & Society*, *28*(1), 40-62.

Xie, Y., & Goyette, K. (2004). *A Demographic Portrait of Asian Americans*. New York, NY: Russell Sage Foundation and Population Reference Bureau.

Xu, J., & Lee, J. C. (2013). The marginalized "model" minority: An empirical examination of the racial triangulation of Asian Americans. *Social Forces*, *91*(4), 1363-1397.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism. *Advances in group processes*, *16*(1), 161-197.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly*, 116-132.

Yamagishi, T., & Mifune, N. (2016). Parochial altruism: Does it explain modern human group psychology?. *Current Opinion in Psychology*, *7*, 39-43.

Yzerbyt, V., & Demoulin, S. (2010). Intergroup Relations. *Handbook of Social Psychology*.

Zaki, J., & Cikara, M. (2015). Addressing empathic failures. Current Directions in Psychological Science, 24, 471-476.

Zárate, M. A., Reyna, C., & Alvarez, M. J. (2019). Cultural inertia, identity, and intergroup dynamics in a changing context. In *Advances in Experimental Social Psychology* (Vol. 59, pp. 175-233). Academic Press.