The Social Neuroscience of Prejudice

David M. Amodio[1,2] and Mina Cikara[3]

[1]New York University, [2]University of Amsterdam, and [3]Harvard University

In press at *Annual Review of Psychology*

An irony of human nature is that while our survival depends on group living, the mere existence of group categories creates prejudice—a preference for one's own group or animus toward another and its members—which leads to discrimination, conflict, and the undermining of society itself (Dovidio & Gaertner, 2010). How do humans learn to favor some groups over others? Why does merely knowing a person's ethnicity or nationality affect how we see them, the emotions we feel toward them, and the way we treat them? Answers to such questions are crucial to our understanding of human social behavior. Although the origins of human prejudices are extraordinarily complex—a multilevel mix of history, geopolitics, social structures, intergroup relations, and social identities—our understanding of how prejudice operates in an individual's mind and behavior has been advanced considerably by the contributions of *social neuroscience* (Amodio, 2014; Kubota et al., 2012).

Social neuroscience is a field of research that probes the connection between the brain and social behavior. It typically does so from two complementary angles. One angle seeks to understand neural functions as they relate to various social processes, with a focus on the operations of specific neural structures, neurotransmitters, or genes. The other seeks to understand psychological processes by applying knowledge about neural function and the tools of cognitive neuroscience. Research on the psychology of prejudice has benefited most from this second approach; by incorporating models and methods of neuroscience, social neuroscientists have made important new discoveries about how humans perceive groups, form and express prejudice, and regulate their intergroup behaviors.

In this article, we present what has been learned so far from the social neuroscience of prejudice. In the following sections, we describe research on how people perceive groups and categorize their members, how prejudice is learned and represented in the mind, how it relates to

judgment, perception, emotion, and behavior, and how its effects may be regulated. Rather than provide an exhaustive list of findings, we take a step back and ask: what has the neuroscience approach revealed, so far, about the psychology of prejudice? In each section, we discuss key social neuroscience findings, consider interpretational challenges and connections with the behavioral literature, and highlight how they advance psychological theories of prejudice.

Social Categorization: The Antecedent of Prejudice

Social interactions are often thought to begin with the perception of a person's face; yet even the first few milliseconds of this perception can be influenced by targets' social categories and the categorization goals of the perceiver. By investigating the processes involved in social categorization with neural assessments, social neuroscience has produced new evidence for top-down effects of group membership on visual processing while detailing the mechanisms through which social categories influence perception. Here, we describe findings from social neuroscience on how we categorize individuals based on visual cues, and how categorization may arise even in the absence of visual cues to group membership.

The Timecourse of Social Categorization. An essential precursor to prejudice is social categorization (Allport, 1954). Although existing behavioral studies suggest that social categorization occurs quickly (Macrae & Bodenhausen, 2000), social neuroscience research has helped illuminate the precise timecourse of social categorization processes (Ito & Bartholow, 2009). In particular, research using event-related potentials (ERPs)—patterns of electroencephalographic (EEG) activity linked to a stimulus (e.g., a face) or action, measured with millisecond resolution—has revealed that social categorization involves multiple distinct processes that unfold rapidly (Amodio et al., 2014).

In an early ERP study of intergroup categorization, Ito and Urland (2003) recorded White participants' EEG signals while they viewed pictures of White and Black male and female faces. Although the participants' task was to classify faces by either their gender or race, ERPs revealed that regardless of the task, neural activity at approximately 120 ms indicated stronger early neural responses to Black than White faces (see also Kubota & Ito, 2007). This initial effect was indicated by the N100 (or N1) ERP component, which reflects early orienting and attention processing in the occipito-parietal and occipito-temporal regions (Clark, Fan, & Hillyard, 1995), perhaps in response to the coarse visual cue of skin tone.

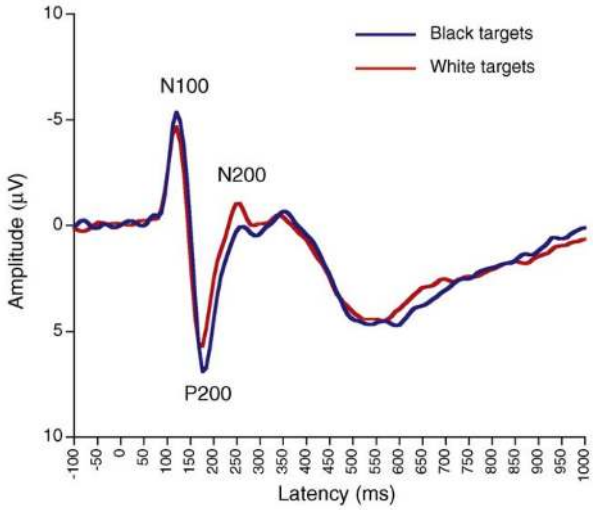

Figure 1. ERP waveforms in response to Black and White faces, viewed by White American participants. Zero ('0') on the x-axis (msec) indicates stimulus onset time. The positive and negative deflections in the waveform represent typical ERP components, named here according to their polarity ('P' for positive deflections and 'N' for negative deflections) and the approximate post-stimulus time (in milliseconds) of their peaks. The N100 and P200 represent early attentional processing of race and are typically larger in response to outgroup members. The N200 is associated with conflict in categorization processing and response formation and is typically larger to racial ingroup members. Negative voltages are plotted above zero on the Y-axis, following electrophysiological convention, although ERP waveforms are sometimes plotted with negative voltages displayed below zero. Adapted from Ito and Urland (2003).

A similar pattern is observed with the P200 (or P2) component, which reflects goal-directed attention and perceptual matching, and peaks at approximately 180-200 ms. The P200 has been shown to differentiate both race, as well as gender. Among White participant perceivers, it is typically larger to Black than White faces (Ito & Urland, 2003). Research with Black participants, in addition to White participants, has replicated this pattern and clarified that it is typically larger to racial outgroup faces rather than Black faces per se (Dickter & Bartholow, 2007; Willadsen-Jensen & Ito, 2008; Volpert-Esmond & Bartholow, 2019). This effect has been observed even when participants are instructed to attend to a target person's gender (Ito & Urland, 2003), to a non-social feature of a face image (Ito & Tomelleri, 2017), or to individuating information (Kubota & Ito, 2017), indicating that the P200 is sensitive to race independent of explicit task instructions. In a study assessing frontal EEG in addition to ERPs in a race priming task, greater left frontal cortical activity—associated with approach motivation and goal activation—predicted larger P200 responses to Black relative to White faces, consistent with the interpretation of the P200 as reflecting goal-directed social categorization (Amodio, 2010). Furthermore, the magnitude of this race-P200 effect has been linked to behavioral expressions of implicit prejudice (Amodio & Swencionis, 2018) and racial bias in a first-person shooter game (Correll et al., 2006).

Depending on the task, these activations may be followed by the N200 (or N2; ~260 ms), such that White American participants typically exhibit larger N200 responses to White than Black faces (Ito & Urland, 2003; Ito & Tomelleri, 2017). Although the psychological significance of this effect is not well understood, the N200 has been associated in other work with response selection and conflict processes because it originates in dorsal anterior cingulate cortex (dACC; Folstein & van Petten, 2008). The typical finding of larger N200 response to

ingroup targets in race categorization tasks may reflect response conflict associated with making an ingroup classification (given the initial tendency to orient to outgroup faces).

Finally, in some tasks (e.g., the classic oddball task), a P300 (or P3; ~450-600 ms) is observed. The P300 has been associated with response evaluation, expectancy violation, and endogenous attention (Bartholow et al., 2001; Ito & Bartholow, 2009) and, in the brain, a distributed set of noradrenergic activations (Nieuwenhuis et al., 2005). Given the late timing of the P300—often following the delivery of a categorization decision in behavior—it may reflect an evaluation of one's response and the updating of task expectations.

Together, ERP studies have begun to characterize the rapid sequence of social categorization processes, beginning as early as 100 ms following face onset and involving stages of category detection, goal-directed attention, classification response selection, and response evaluation (Figure 2).

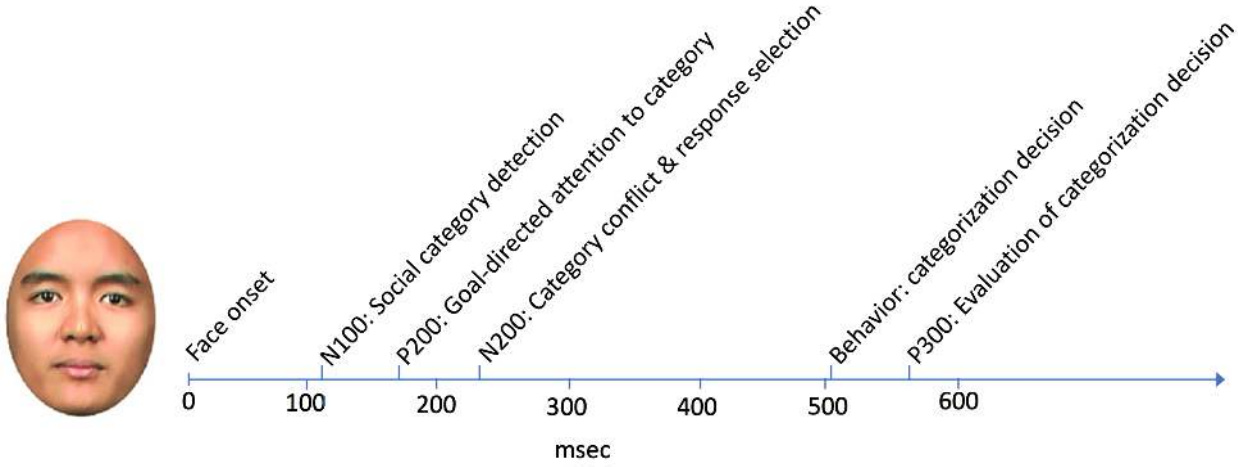

Figure 2. ERP component responses to a face and their putative functions in race categorization tasks, showing the typical timing and interpretation of each component, in addition to the timing of a behavioral categorization response.

Further support for the early detection and categorization of race is suggested by race effects in primary visual cortex (V1), observed in functional magnetic resonance imaging (fMRI)

studies. Using multi-voxel pattern analysis (MVPA), an analytic technique that uses patterns of brain activity to differentiate between mental states or representations, these studies found that patterns of activity in this region could decode the race of a face (Brosch et al., 2013; Gilbert et al., 2012). In another study, MVPA revealed that an individual's arbitrary group membership, independent of race, was also able to be decoded in V1 (Ratner et al., 2013). These fMRI results corroborate the early categorization effects seen in ERPs by showing race and arbitrary group detection in V1—the anatomical starting point of the cortical visual stream.

In some cases, a person may be perceived according to multiple social categories (e.g., race and gender). In this context, fMRI research has begun to reveal the complex and dynamic interplay of top-down and bottom-processes involved in social perception (Freeman & Johnson, 2016). For example, this research has shown that overlap in a perceiver's mental representation of two social categories (e.g., race and gender) correlates with the degree to which neural patterns linked to each category are activated in the fusiform cortex when viewing a face (Stolier & Freeman, 2016). These data suggest that as a face is being encoded, preexisting cognitive representations of social categories in the anterior temporal lobe and orbital frontal cortex converge with visual inputs in the fusiform cortex through a rapid iterative process to shape the perception of social category membership. When a single-category decision is required, ambiguity in these representations is resolved with input from the dACC (Stolier & Freeman, 2017), which is broadly involved in the detection of conflict and allocation of control (Shenhav et al., 2013). Other research has linked individual differences in neural patterns associated with racial categorization to prejudice (e.g., biased altruism intentions; Zhou et al., 2020). Together, these findings begin to elucidate the neural and psychological processes involved in the initial perception and social categorization of a person's face.

Categorization in the absence of visual cues to group membership. In everyday life, social categorization is highly context dependent (Turner et al., 1994), with particular category distinctions emerging over the course of a perceiver's experience as their goals and situations change. How do people distinguish ingroup from outgroup members in dynamic environments with other agents and their respective, intersecting group memberships? By some accounts, categorizing people by specific social categories is a byproduct of adaptations that evolved for detecting more general coalitions (Sidanius & Pratto, 2012; Pietraszewski, Cosmides, & Tooby, 2014). Such accounts suggest that humans need a flexible, common neural code for learning about and representing 'ingroup' and 'outgroup' 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? And, more importantly, what would be the primary structure of the code (e.g., ingroup vs. everyone else, threatening outgroup vs. everyone else, distinct codes for ingroup, neutral outgroups, and threatening outgroups)?

To adjudicate among these competing categorization structures, one fMRI study used MVPA to test whether participants' neural responses associated with thinking about political partisans (Democrats v. Republicans) could be used to successfully decode whether they were thinking about teammates as opposed to competitors created in the lab (Rattlers v. Eagles; Cikara et al., 2017). Only two regions were associated with representing the higher-order concepts of "us" vs. "them" across both political and lab-based groups: the dorsal ACC/middle cingulate cortex and the anterior insula (AI). The dACC (referenced above) and AI have been posited as hubs in a 'salience network,' which focuses attention on the most relevant internal and external stimuli (both social and nonsocial) 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 ingroup is consistent with the hypothesis that salience—specifically functional significance or evaluation (e.g., will this person help me or not?)—is the primary dimension distinguishing representations of us and them (Fiske, 2018; see however, Koch et al., 2016). Furthermore, this analysis revealed the structure of this neural code: classification accuracy across categories was driven predominantly by the correct categorization of ingroup targets, consistent with theories indicating ingroup identity and preference are more central than outgroup processing in group perception and cognition (Balliet et al., 2014; Brewer, 1999).

But how do people resolve the challenge of categorization in the absence of labels or visual cues to group membership? One possibility is that they simply substitute judgments of similarity to one's self on relevant features (e.g., how did you vote in the last election?). In line with this proposition, neuroimaging studies report that a ventral region of medial prefrontal cortex (vmPFC)—which has been associated with thinking about one's own, as well as similar others' traits, mental states, and characteristics (Denny et al., 2012; Jenkins & Mitchell, 2011)—is also more engaged when people categorize ingroup relative to outgroup members (Molenberghs & Morrison, 2012; Morrison et al., 2012).

However, in addition to relying on similarity as an input, people's inferences about social group dynamics 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 do I get along with Susan? With Doug? How do they get along?"). This approach allows perceivers to infer social group structure (i.e., clusters over individuals; Gershman & Cikara, 2020).

In a series of behavioral experiments framed as learning about strangers' political issue positions, 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. Specifically, participants favored Agent B over 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 (see Figure 3; Lau et al., 2018). In a companion fMRI study (Lau, Gershman, & Cikara, 2020), trial-by-trial estimates of similarity between participants and each individual agent recruited vmPFC and pregenual anterior cingulate, in line with previous work. By contrast, latent social group structure-based estimates recruited right AI (which overlapped with a region identified by a *non-social* structure learning task; Tomov et al., 2018), suggesting that rAI supports domain-general structure representation. Most interesting, however, was that 'social group structure' neural signals further explained ally-choice behavior whereas 'inter-agent similarity' signals did not. This suggests that people base their identification of their ingroup more on the structure of the group as a whole than on our own similarity with individual group members.

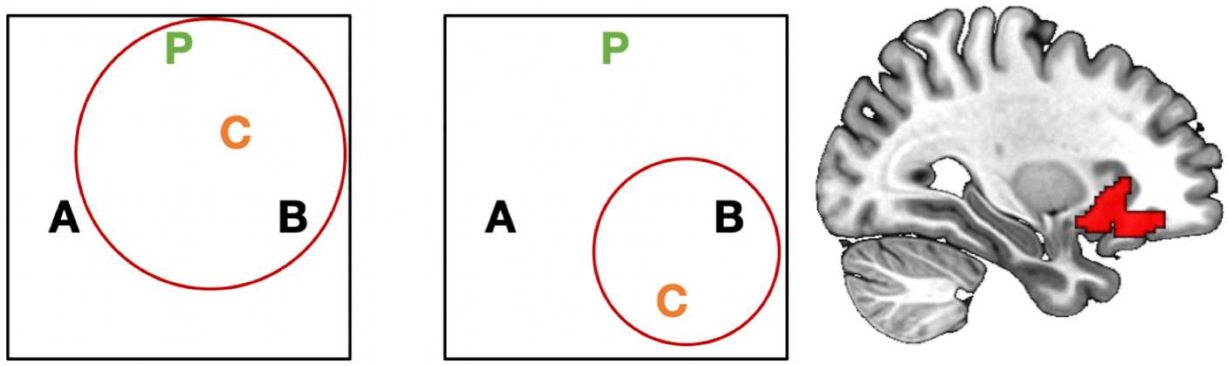

Figure 3. *Left:* Schematic representation of different social structures as a function of Agent C (Lau et al., 2018). Distance is a proxy for similarity. In both panels A and B are equally similar to you, but in the left panel C's placement creates a group that includes both you and B (which increases preference for B relative to A), whereas in the right panel C's placement puts B in a group that does not include you. *Right:* Results from whole-brain contrast (FWE-corrected $p < 0.05$) of latent structure learning parametric modulators: right anterior insula ($x = 30$).

Summary: Social Categorization. Social neuroscience research has significantly advanced our understanding of the social categorization process by delineating its timing and sub-processes in ERP studies and, in recent fMRI research, by addressing the neural and psychological processes through which categorization unfolds in more complex, intersectional social environments. In line with theorizing of intergroup relations on the basis of functional relations (Fiske, 2018; Koch et al., 2016), these results suggest that generalized group concepts rely on domain-general circuitry associated with latent structure learning and the encoding of stimuli's functional significance.

How is prejudice learned, represented, and activated?

One of the most intriguing findings in intergroup psychology is that prejudiced responses are activated automatically upon encountering a group-based cue—an effect that connects the perception of a group member to the activation of the perceiver's prejudice (Devine, 1989; Fazio et al., 1995). Although this effect has been widely replicated, many questions remain; for example, how are these automatic associations learned? How are they represented in the mind? And how do they affect behavior? Social neuroscience research has shed considerable new light on these issues by integrating theory and methods from neuroscience, particularly as they relate to learning and memory, to address questions about prejudice.

Although the traditional view in social cognition assumes that intergroup associations are formed and represented in a single semantic network, we now know that human learning and memory involves multiple interacting neurocognitive systems (Squire & Zola, 1996; Poldrack & Foerde, 2008). A consideration of multiple memory systems is important because it suggests that multiple kinds of information are encoded, beyond semantic knowledge, and that these different

kinds of information are expressed in particular channels of behavior. These systems include memory processes addressed in traditional prejudice research, such as semantic (or conceptual) knowledge and associations, as well as others that have only recently been applied to human social cognition and prejudice, such as Pavlovian and instrumental learning (Amodio, 2019; Amodio & Ratner, 2011a). A sample of these learning and memory systems is shown in Figure 4, along with their respective neural substrates and putative channels of expression. In this section, we describe advances in our understanding of how intergroup bias is learned and represented in the mind, based on contemporary neuroscience models of learning and memory, and discuss their implications for how biases may be activated and expressed in behavior.

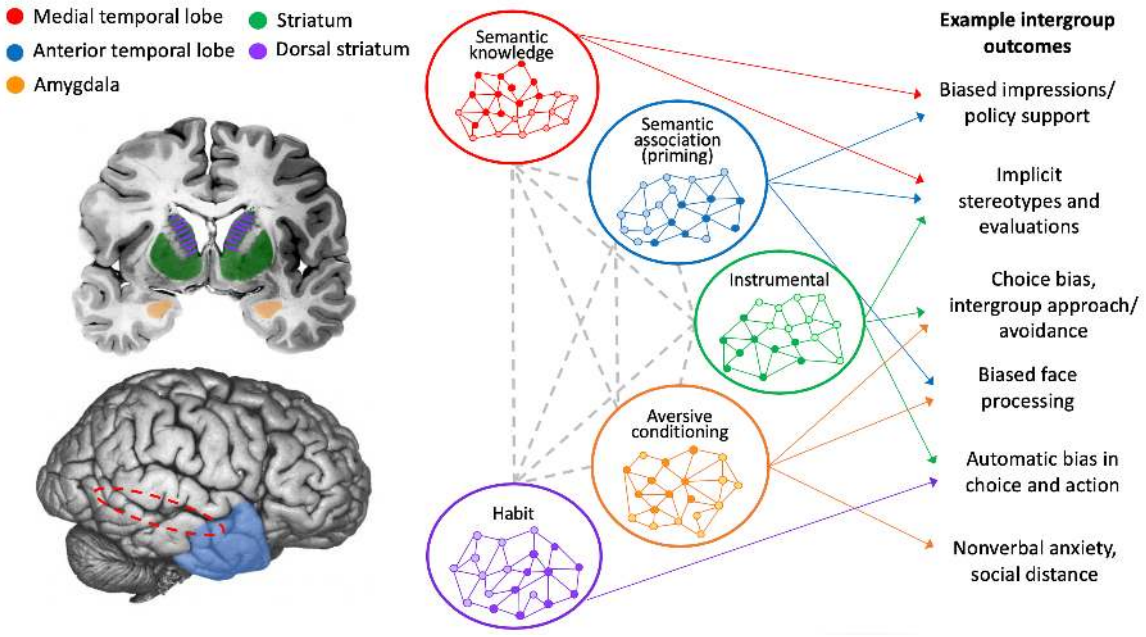

Figure 4. A model of the learning and memory systems through which forms of intergroup bias are acquired and represented, with illustrations of their putative neural substrates and examples of their respective intergroup outcomes. Adapted from Amodio (2019).

An affective basis of implicit prejudice? The role of Pavlovian aversive conditioning. An enduring, yet complicated, idea in the social neuroscience of prejudice is that the amygdala

underlies implicit prejudice. This idea is complicated because evidence for the amygdala's role in prejudice is mixed, yet the notion that Pavlovian aversive conditioning—learning to fear a neutral stimulus—could contribute to bias formation remains plausible. The amygdala is a small structure located bilaterally within the temporal poles. Given its critical role in Pavlovian aversive conditioning, it was initially regarded as the neural center of learned fear in both animals and humans (LeDoux & Hoffman, 2018). Although this "fear center" interpretation has since been revised and elaborated (e.g., Holland & Gallagher, 1999; LeDoux, 2012), the idea that the amygdala, and its role in Pavlovian aversive conditioning, could underlie implicit bias remains intuitive and intriguing to prejudice researchers.

Consider the amygdala's neural circuitry: Signal of a learned threat can travel from its initial sensation, in the retina or cochlea, to the amygdala via a single synapse, such that the amygdala can initiate a defensive response within approximately 100 ms (LeDoux & Hoffman, 2018). Perceptual information enters the amygdala via the lateral nucleus and, if associated with a learned threat, activates the central nucleus, which in turn initiates freezing and heightened vigilance (e.g., potentiated startle) in preparation for fight or flight. This rapid response occurs while more elaborative processing continues in other neural regions—a pattern resembling dual-process accounts of prejudice in which an automatic response proceeds before a more deliberative response can take over (e.g., Devine, 1989). These characteristics have several implications for theories of prejudice.

First, research on the amygdala and aversive conditioning suggests a distinct affective basis for acquiring prejudice, as well as a plausible mechanism to explain the rapid, nonconscious, and unintentional negative responses to racial outgroup members that characterize automatic prejudice. Like most other animals, human acquire fear-conditioned responses to

stimuli (Delgado, Olsson, & Phelps, 2006), including humans faces (Öhman & Dimberg, 1978), and thus, in theory, this mechanism could also support learned aversions to groups. Some research has attempted to demonstrate a Pavlovian basis of prejudice using prepared fear or reversal learning paradigms (Dunsmoor et al., 2016; Olsson et al., 2005), but these results have been inconclusive regarding a prepared fear to Black faces (among White Participants) or have failed to replicate (Mallan et al., 2009; Molapour et al., 2015; Navarrete et al., 2009; Navarette et al., 2012). To our knowledge, research has not yet directly tested the hypothesis that social prejudice can be formed through Pavlovian aversive conditioning.

Second, an aversive conditioning model of prejudice is useful because it predicts a particular pattern of behavior in human intergroup interactions—that of freezing, anxiety, vigilance, and avoidance. Similar behaviors have been observed in social psychological studies of intergroup interactions; for example, anti-Black prejudice in White participants has been associated with adopting greater physical distance from Black partners (Amodio & Devine, 2006; McConnell & Liebold, 2001), heightened vigilance (Richeson & Trawalter, 2008), nonverbal signs of anxiety (Dovidio et al., 2002; Fazio et al., 1995), and physiological arousal (Amodio, 2009; Trawalter et al., 2012). It further explains why intergroup anxiety amplifies implicit prejudice but not implicit stereotyping (Amodio & Hamilton, 2012). Hence, an aversive conditioning mechanism of bias, while novel to psychological theories of prejudice, helps to explain a broader range of prejudiced behaviors.

Third, and more broadly, social neuroscience research positing an aversive conditioning component of prejudice sparked a paradigm shift in social cognitive models of prejudice. Whereas prior theories assumed that prejudice emerges from a single cognitive network of semantic concepts (i.e., stereotype knowledge), conditioned fear (a) involves threat associations,

formed through highly-arousing aversive experiences and (b) is expressed primarily in behavior and autonomic arousal. Hence, this research revealed a second mechanism for social learning and prejudice and, by linking the study of prejudice to broader models of learning and memory, pointed to additional mechanisms of social learning and prejudice that had yet to be studied (Amodio & Ratner, 2011a; March et al., 2018).

It is notable, however, that despite the existence of Pavlovian aversive conditioning in humans and its likely role in nonverbal and affective expressions of prejudice, neuroimaging evidence for a stronger amygdala response to racial outgroup members has been mixed, at best (Checkroud et al., 2014). Indeed, most fMRI studies of race perception have not observed a difference in amygdala response to viewing racial outgroup compared with ingroup members (e.g., Beer et al., 2008; Gilbert et al., 2012; Golby et al., 2001; Knutson et al., 2007; Mattan et al., 2018; Phelps et al., 2000; Richeson et al., 2003; Ronquillo et al., 2005; Stanley et al., 2012; Telzer et al., 2013; Van Bavel et al., 2008, 2011). Of those that did, race effects were observed under specific conditions: for example, when Black and White faces were presented very briefly (Cunningham et al., 2004), when participants made superficial rather than individuating judgments (Wheeler & Fiske, 2005), or when the target face's gaze was direct but not averted (Richeson et al., 2008). Other research found that the amygdala effect—greater to Black than White faces—was stronger among African American participants than White participants (Lieberman et al., 2005). Notwithstanding limitations common to early fMRI studies (e.g., small samples, less stringent corrections for multiple comparisons), these instances of positive findings, in which amygdala effects were observed under some conditions but not others, suggest a more complex account of the amygdala's role in race perception.

Research using the startle eyeblink method to assess the amygdala response to racial outgroups has added to our understanding of its role in prejudice. These studies suggest that the amygdala primarily guides attention to race, based on its motivational relevance, especially in situations of threat or anxiety. This perspective stems from the method's amenability to larger sample sizes and more varied experimental designs, compared with fMRI, as well as its historical roots in research on attention and motivation (Filion et al., 1998). For example, an early study of White participants found greater startle response to Black faces than to both White and Asian faces (Amodio et al., 2003). Although this finding was initially interpreted as revealing an amygdala substrate for prejudice, further analysis suggested that this effect was primarily associated with participants' anxiety about appearing prejudiced to others (i.e., their external motivation to respond without prejudice), even among people with egalitarian attitudes. Subsequent startle eyeblink and fMRI studies similarly found that amygdala responds not to race per se, but to social goals and task strategies (Brown et al., 2006; Mattan et al., 2018; Vanman et al., 2013; Van Bavel et al., 2008; Wheeler & Fiske, 2005). That is, these findings suggest that the amygdala response to racial outgroup members often reflects attention driven by social goals and concerns, rather than the direct threat of an outgroup member (Amodio, 2014; Checkroud et al., 2014). Moreover, social concerns about appearing prejudiced have been linked to implicit prejudice (Devine et al., 2002); this link may explain why the amygdala response to race has been found to correlate with implicit prejudice in some work (e.g., Phelps et al., 2000).

In summary, Pavlovian aversive conditioning likely contributes to a specific aspect of prejudice—one that operates automatically, is associated with negative affect, and is expressed in nonverbal behaviors such as freezing and social distancing. However, despite early excitement about the possibility that the amygdala underlies implicit prejudice, this idea has not been

supported by the fMRI literature. Instead, amygdala activations in intergroup contexts appear to reflect attention to relevant group cues, as determined by one's social motivations and goals, or one's anxiety about appearing prejudiced. Nevertheless, the role of the amygdala in prejudice formation remains plausible and ripe for study, as a Pavlovian learning process provides the best account of some forms of intergroup behavior.

Stereotypes and conceptual evaluations: The role of semantic memory. Stereotypes represent the conceptual attributes linked to a particular group, as defined within a particular culture or society. Stereotyping involves the encoding and storage of group-based concepts, the selection and activation of these concepts into working memory, and their application in judgments and behaviors (Fiske, 1998). As such, stereotyping involves cortical structures that support more general forms of semantic memory, object memory, retrieval, and conceptual activation, such as the temporal lobes and inferior frontal gyrus (IFG; Martin, 2007), as well as regions involved in impression formation, such as the medial prefrontal cortex (mPFC; Amodio & Frith, 2006). Social knowledge—about people and groups—has been specifically linked to anterior temporal lobe (ATL), including the temporal pole (Olson et al., 2013; Zahn et al., 2007). Hence, stereotypes and conceptual evaluations—to the extent they represent a social form of semantic processing—should also be associated with activity in these regions.

In an fMRI study of racial stereotypes, Gilbert et al. (2012) used MVPA to dissociate neural activity representing judgments of Black and White individuals on the basis of either stereotype-associated traits (athleticism) or evaluations (potential friendship). Race-based differences in stereotype trait judgments were represented in the mPFC, similar to observations of gender stereotype judgments (Contreras et al., 2012; Quadflieg et al., 2009), whereas evaluative judgments were represented in OFC (Gilbert et al., 2012). To probe stored

representations of stereotypes and evaluations, the authors looked for regions in which multi-voxel patterns could reliably predict participants' scores on independent implicit association test (IAT) measures of racial stereotyping and evaluation, respectively. They found one region that accurately represented both implicit stereotyping and implicit evaluation: the ATL. That is, when subjects made trait judgements, stereotyping IAT scores were associated with one pattern of ATL activity; when they made evaluative judgements, evaluative IAT scores were associated with a different pattern within the same region. These findings support a semantic memory basis for implicit bias rooted in conceptual associations, including both stereotypes and evaluations.

Consistent with an ATL substrate of stereotype representation, Spiers et al. (2017) observed that the formation of racial stereotypes, acquired as participants read descriptions of outgroup members' negative behaviors, was tracked uniquely by activity in the temporal poles. In other research, disruption of ATL activity via transcranial magnetic stimulation (TMS) attenuated the behavioral expression of implicit gender stereotype associations (Gallete et al., 2011). Furthermore, ERP studies have linked stereotype processing to the N400 ERP component (e.g., White et al., 2009), a neural signal originating from the temporal lobe that is associated with language and semantic memory processes and occurs ~400 milliseconds following word presentation (Kutas & Federmeier, 2011). When judging a novel group member, group stereotypes represented in the ATL may influence one's impression via signals to the mPFC (Amodio, 2014), consistent with anatomical connections between these regions (Olson et al., 2013). Hence, while the neural basis of stereotyping remains understudied, existing research consistently identifies the ATL as supporting the representation of social stereotypes and, through connectivity with the mPFC, the application of stereotypes in impression formation.

Prejudice formation through social interaction: The role of instrumental learning.

Ironically, most psychological research on impression formation concerns indirect experiences of others—in lab studies we learn about others by reading descriptions, observing behaviors, or applying stereotypes. Yet much of real-life social behavior involves direct interaction, and thus a current major goal of social cognition research is to understand how we form impressions of people and their groups through social exchange. Recent social neuroscience findings suggest this form of direct interaction-based social cognition may be rooted in instrumental learning—a mode of feedback-based reward reinforcement associated with activity of the striatum (Hackel et al., 2015). The striatum, which comprises the caudate, nucleus accumbens, and putamen, supports the learning and representation of reward value and, through its connectivity with the PFC and motor cortex, guides choice and goal-directed action (O'Doherty et al., 2017).

Although social psychologists have long hypothesized a role for instrumental learning in attitudes and social behavior (e.g., Breckler, 1984), this idea has only recently been tested using contemporary reinforcement learning paradigms and computational modeling (Behrens et al., 2009; Hackel & Amodio, 2018). Behavioral studies confirm that people incrementally update their attitudes about both persons (Hackel et al., 2019) and groups (Kurdi et al., 2019) in a manner predicted by reinforcement models. Convergent fMRI research has linked this process to the striatum (Hackel et al., 2015). Human learners can similarly form and update trait-like inferences in response to feedback (Hackel et al., 2015, 2020)—a process supported by the striatum in combination with regions often implicated in social cognition (e.g., rTPJ, precuneus, intraparietal lobule). These findings suggest that instrumental learning may support both an action-based form of social attitude as well as the formation of conceptual trait impressions.

In the context of prejudice, instrumental learning represents the formation of reward associations through repeated action and feedback, for example, through the process of approaching an ingroup or outgroup member and encoding their response. Instrumental associations should be more directly linked to action, given their learning mode and underlying neural circuitry, relative to semantic or Pavlovian associations, and thus instrumental forms of prejudice may be most strongly expressed in behavior (Amodio & Ratner, 2011a). Unlike semantic learning, which pertains to specific conceptual associations, instrumental learning represents probabilistic reward associations and does not require awareness for its learning or expression (Knowlton et al., 1996). For this reason, a model of instrumental prejudice may help to understand aspects of implicit prejudice—particularly those expressed via action, as opposed to those observed in word associations. Finally, instrumental associations are malleable, fluctuating according to the reward history of a social target, in contrast to Pavlovian associations, which are difficult to alter (LeDoux & Hofmann, 2018). Thus, manipulations known to change instrumental reward associations may inform new interventions for how to reduce this aspect of prejudice. Predictions such as these, based on the emerging literature on instrumental learning in social cognition, are currently guiding a new wave of research on the social neuroscience on prejudice.

Habits: A basis for automatic prejudice? Automatic prejudices are often likened to habits; they appear to emerge from repeated negative experiences with outgroup members, unfold without intention, and resist change (Devine, 1989). While an intuitive analogy, is there evidence that prejudice can operate like a habit?

Habits typically emerge from instrumental learning—responses that begin as a goal-directed reward-driven actions which, over time and repetitions, become routinized as automatic

responses that persist irrespective of reward (Wood & Neal, 2017; Robbins & Costa, 2017). Whereas goal-directed instrumental learning is primarily associated with the ventral striatum, habit-driven responses have been linked to the dorsal striatum (Foerde, 2018).

Although social neuroscience has yet to investigate the role of habit in prejudice, behavioral research suggests that a habit-like process, such as model-free learning, can underlie social attitudes toward both persons and groups (Hackel et al., 2019; Kurdi et al., 2019). These findings suggest that habits may indeed play a role in prejudice. However, unlike the "habit" analogy for automatic stereotyping, a habit component of prejudice would most likely be expressed in action and choice, given its roots in instrumental learning. While further research is needed, a consideration of habits as a mechanism for prejudice promises to inform our understanding of how implicit bias is expressed and potentially reduced.

Summary: The social neuroscience of prejudice formation and representation. A major contribution of social neuroscience research on prejudice has been to link different aspects of prejudice—stereotypes, affective bias, and discriminatory actions—to neurocognitive models of learning and memory. It reveals that intergroup bias, and implicit bias in particular, is not one phenomenon, but a set of different processes that may be formed, represented in the mind, expressed in behavior, and potentially changed via distinct interventions.

Effects of prejudice on perception, emotion, and decision-making

Once categorization has occurred and prejudice is activated, it modulates other psychological processes—what we see, how we feel, and how we form judgments—all of which can influence behavior. In this section, we review discoveries from social neuroscience on the effects of prejudice on face perception, intergroup emotion, and decision-making.

Face perception. Since the "New Look" proposal that motivation influences object perception, prejudice researchers have considered the possibility that prejudice shapes how we see ingroup and outgroup members (Kawakami et al., 2017). Social neuroscience has advanced this line of inquiry by introducing methods from vision neuroscience to complement behavioral methods that, on their own, cannot easily discern changes in perception from changes in a person's *judgment* of their perception. In doing so, this approach has produced new and more rigorous evidence for the effects of prejudice on early face processing while elucidating the mechanisms through which top-down social factors influence social perception. In contrast to the categorization research we discussed above, what follows is a review of work that seeks to understand more specifically how prejudice-biased perception gives rise to discriminatory phenomena (e.g., race-based misidentification in line-ups).

Humans are expert face perceivers, and the capacity to identify a human face, encode its features, track its orientation, and recognize its identity is supported by an extensive network of neural regions that include the fusiform, occipital cortex, and temporal lobe (Duchaine & Yovel, 2015). An initial stage of face perception is the configural encoding of a stimulus as a face—that is, determining that the arrangement of an object's features matches the canonical configuration of a human face: two eyes above a nose, above a mouth. Simultaneously, the brain encodes specific facial features, although configural processing is typically prioritized. Configural face processing is associated primarily with the fusiform gyrus, whereas featural processing occurs in temporo-occipital cortex (Duchaine & Yovel, 2015).

In an early fMRI study of the *own-race bias* effect, whereby ingroup faces are recognized better than outgroup faces, Golby et al. (2001) observed greater activity in the fusiform gyrus when White participants' viewed ingroup than outgroup faces, and this neural pattern predicted

better memory for ingroup faces. This finding revealed greater configural encoding of ingroup than outgroup faces—a difference in the early perceptual encoding of an image as a human face. More recent work, which examined the effect of race on a phenomenon called "repetition suppression," suggests that prioritized ingroup processing in the fusiform contributes to the *outgroup homogeneity effect*, which, similar to the own-race bias effect, refers to people's tendency to view outgroup members as less distinguishable than ingroup members (Hughes et al., 2019; Reggev et al., 2020).

Most studies examining race effects on face perception have used an ERP approach, with a focus on the face selective N170 component—a neural signal associated with the initial configural encoding of a face, which is generated in the fusiform and temporo-occipital cortices and occurs at just ~170 ms after face onset. Early effects of race on the N170 appeared mixed—some found larger responses to racial ingroups (Ito & Urland, 2005; Feng et al., 2011), others to racial outgroups (Walker et al., 2008), and many others found no differences (e.g., He et al., 2009; Wiese et al., 2009; Caldara et al., 2003). However, more recent research has clarified that group effects on face encoding depend on a perceiver's task goals and social motivations (Ofan et al., 2011; Senholzi & Ito, 2013). When race is relevant to one's goal, configural processing of goal-relevant group members is enhanced; when race is not relevant, faces of both groups are processed similarly. For example, when a Black face represents a threat cue (e.g., because a participants' group dominance motives were activated, or because the participant was worried about appearing prejudiced to others; Ofan et al., 2014; Schmid & Amodio, 2017), the N170 may be larger to Black than White faces. By contrast, when a White participant is motivated to discount or stereotype outgroup members, their N170 response may be smaller to Black than White faces (Schmid & Amodio, 2017).

Several factors have now been shown to influence the effect of race on configural face encoding, such as categorization goals (Ito & Urland, 2005), social power (Schmid & Amodio, 2017), economic scarcity (Krosch & Amodio, 2019), implicit prejudice (Ofan et al., 2011), intergroup anxiety (Ofan et al., 2014), perceiver race (Vizioli et al., 2010), group identity (Scheepers et al., 2013), and intergroup contact (Walker et al., 2008). Such effects have been found among people of many different nationalities, including Canadian, Chilean, Chinese, Israeli, Japanese, Korean, and Swiss, and their relevant ethnic outgroups (e.g., Caldara et al., 2003; Ibanez et al., 2010). Increased configural processing, as indicated by the N170 or fMRI measures of fusiform activity, has also been observed for novel (Van Bavel et al., 2011) and minimal (Ratner & Amodio, 2013) ingroup members, university ingroup members (Cassidy et al., 2014), and sex-typical faces relative to sex-atypical faces (Freeman et al., 2010). In some studies, the N170 to racial outgroups was also delayed (Ofan et al., 2011; Stahl et al., 2008; Weise et al, 2008; Zheng & Segalowitz, 2014)—a pattern is consistent with a shift to feature-based processing as a result of impaired configural processing (Rossion et al., 2000). Collectively, this research demonstrates an effect of intergroup bias on the earliest stages of face processing that, under certain conditions, may impede a perceivers' ability to process outgroup faces the same way as ingroup faces—an effect that has been dubbed *perceptual dehumanization* (Fincher & Tetlock, 2016; Kawakami et al., 2017) and linked to outgroup homogeneity effects (Hughes et al., 2019).

Most important, these race effects on configural encoding may function to justify and promote discriminatory behavior (Krosch & Amodio, 2019). In complementary ERP and fMRI studies, White participants determined how much money each of a set of White and Black individuals deserved. Participants exhibited a selective delay in the N170 (using EEG) and

reduction in fusiform activity (using fMRI) to Black, compared with White, faces that emerged only under conditions of perceived economic scarcity (Figure 5). Moreover, in both studies, the magnitude of this encoding deficit was associated with the degree of anti-Black disparity in participants' money allocations. These findings are consistent with the idea that intergroup prejudice (e.g., induced by scarcity) can lead perceivers to view outgroup members in a way that facilitates harmful behavior (Fincher & Tetlock, 2016; Rai et al., 2017; Zhou et al., 2020).

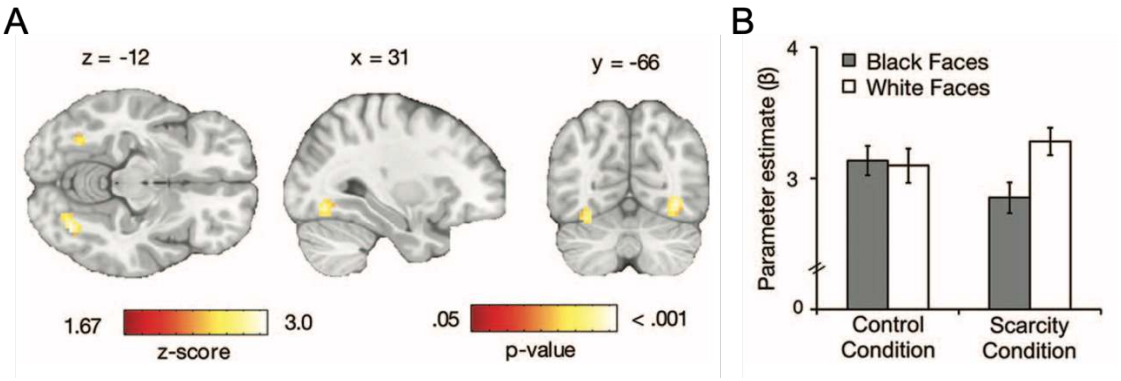

Figure 5. Face-selective activity in the fusiform cortex (panel A) was reduced among White participants when they viewed Black faces, relative to White faces, in a condition of perceived scarcity (panel B)—a pattern associated with racial bias in participants monetary allocations to Black and White recipients. Adapted from Krosch and Amodio (2019).

Together, these studies reveal that prejudice and intergroup dynamics can indeed shape the earliest stages of face processing, and that they do so flexibly and in a goal-consistent manner. Moreover, by identifying specific factors that affect early social perception (e.g., prejudice, power, scarcity), this work suggests contexts in which the effects of prejudice on perception may be modulated, and thus potentially reduced.

Emotion. A central goal of prejudice research is to inform our understanding of discrimination and intergroup behavior. Although prejudice is typically measured in terms of an attitude—that is, on a single dimension of valence, ranging from negative to positive—attitudes are often not fine-grained enough to predict specific behaviors; for example, when do negative

attitudes predict neglect, as opposed to fear or attack (Fiske, 2018)? To understand the specific behaviors associated with prejudice, a more nuanced analysis of discrete intergroup emotions is needed (Neuberg & Schaller, 2016; Mackie & Smith, 2018).

A distinctive feature of intergroup emotions is that they may conflict with the emotional responses people feel in interpersonal contexts. In other words, in intergroup contexts, people's emotional responses may shift to reflect the priorities and interests of the group instead of the individual (Mackie & Smith, 2018). Nowhere is this pattern better characterized than in the domain of how we feel in response to ingroup versus outgroup members' suffering. The social neuroscience of intergroup empathy has illuminated that there are distinct pathways that contribute to ingroup help/outgroup neglect versus outgroup harm (Vollberg & Cikara, 2018).

Empathy is a multi-faceted construct, comprising both cognitive and affective components that reflect our reactions to others' experiences and feelings. Understanding a target's experience in the absence of any concomitant affect has been associated with a distributed set of brain regions including mPFC, temporoparietal junction, temporal pole, and precuneus—regions involved in the representation of trait impressions, perspective-taking, person knowledge, and self-awareness, respectively (Amodio & Frith, 2006; Saxe, 2010; Olson et al., 2013). Experiencing an emotion in reaction to someone else's emotion, on the other hand, is typically associated with engagement of dACC and AI (Zaki & Ochsner, 2012; Lamm et al., 2019). Because the AI and dACC are associated with both the first-hand experience of pain and empathy for others, early theories posited that the affective components of empathy were the product of simulating others' pain (Hein & Singer, 2008). However, both regions are involved in a range of functions, including the detection of cognitive conflict, tracking of value, and salience (see also section on categorization above). Therefore, more recent formulations posit dACC and

AI consistently correlate with empathy due to their general function of encoding salient cues and value (Decety, 2011).

While there remains ambiguity surrounding the precise functions of these regions in the experience of empathy, there is relatively greater consensus surrounding the phenomenon of *intergroup empathy bias*. Dozens of physiological, fMRI, and EEG studies indicate that people are less likely to empathize with others when they are socially distant, such as when they belong to different racial or national groups (Cikara et al, 2011; Cikara & Van Bavel, 2014; Han, 2018). For example, participants in an fMRI study exhibited greater dACC engagement when watching members of their racial ingroup (Caucasian or Chinese) relative to the outgroup being pricked by a needle (Xu et al., 2009). This dACC and AI bias pattern has replicated across cultures, including Chinese (Sheng et al., 2014), Australian (Contreras-Huertas et al., 2013), and European (Azevedo et al,. 2013) participants, and across group contexts, including sports fans (Cikara et al., 2011; Hein et al., 2010).

It is notable, however, that findings from at least two studies diverged from this pattern. In the first case, participants who viewed images of same-race and other-race targets suffering in the aftermath of Hurricane Katrina exhibited similar degrees of dACC and AI activation across both conditions (Mathur et al., 2010). Similarly, Arabs and Israelis exhibited equivalent dACC and AI responses to stories of ingroup and outgroup pain (Bruneau et al., 2012). These patterns may also be moderated by the majority/minority status or power of the groups under inquiry (e.g., Black vs. White participants in South Africa viewing Black and White targets' suffering; Fourie et al., 2017). Future work is tasked with determining whether these discrepancies are due to differences in samples, stimulus sets, or statistical power.

Similar patterns have been documented via reduced motor resonance—activation of an observer's motor system, attuned to the perceived movement of another—with outgroup relative to ingroup targets (Avenanti et al., 2010; Fini et al., 2013; Gutsell & Inzlicht, 2012). For example, watching ingroup members as opposed to outgroup members receive an injection resulted in increased event-related desynchronization of beta rhythms in sensorimotor cortex, which the authors interpreted as greater resonance with ingroup pain (Riečanský et al., 2015; see also Levy et al., 2016).

However, lapses in empathy alone cannot explain overt intergroup conflict. After all, the absence of empathy is merely apathy, which is generally not a strong predictor of aggressive behavior. Thus, a growing body of work has focused on understanding the conditions under which people experience the exact opposite of empathy—specifically, pleasure in response to others' misfortunes (*Schadenfreude)*. People are least likely to experience empathy and most likely to experience Schadenfreude in intergroup contexts when they see outgroups as both competitive with their own interests and high-status: not only are "their" goals at odds with "ours;" they also pose a legitimate threat (Cikara, 2015; Harris et al., 2008). In an fMRI study testing the link between Schadenfreude and harm (Cikara et al., 2011), Red Sox and Yankees fans reported how much they felt pleasure, anger, and pain after watching baseball plays in which their team and their rival scored or failed. Not surprisingly, participants reported feeling pleasure when players on their own team succeeded and a rival team failed, even against the Orioles (a relatively less competitive team in the same league). Pleasurable baseball plays, including rivals failing to score against the Orioles (the pure Schadenfreude condition), activated responses in the ventral striatum (VS), a region associated with learning from *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). Note also that no such correlation emerged with dACC or AI (mirroring the absence of a relationship between reduced empathy and aggression; Vachon, Lynam, & Johnson, 2014). In a related fMRI study, soccer fans exhibited VS activity when watching a rival team's *fan*—someone who is merely affiliated with the rival team—receive a painful electric shock. Increased VS in this context was correlated with a decreased willingness to help the rival fan (Hein et al., 2010).

The unique association of outgroup harm with activity in the VS is notable because there are several regions in the brain associated with the registration of pleasure (including AI, vmPFC and medial orbitofrontal cortex). VS, however, is associated with reward prediction errors for the purposes of planning future behavior. According to one model, the capacity for intergroup aggression may have developed, in part, by appropriating basic reinforcement-learning processes and associated neural circuitry—including VS—to overcome harm aversion (Cikara, 2015). As such, the repeated experience of Schadenfreude in response to outgroup suffering may be the slippery slope that slowly transforms unthinkable actions to acceptable ones.

As we have emphasized throughout this article, these emotional responses are malleable and context-dependent. If the nature of one's relationship with an outgroup member changes, their degree of empathy follows. For example, participants expressed greater empathy toward an outgroup member who volunteered to receive electric shocks in order to spare the participant, in comparison to an ingroup member who did the same. Specifically, greater responses in AI associated with *receiving* help from an outgroup member predicted significantly greater AI activation in response to seeing other outgroup members in pain (relative to a baseline, before they received help; Hein et al., 2016).

Social neuroscience research has also expanded our understanding of guilt, which, in response to one's intergroup transgression, is a powerful elicitor of self-regulation and prosocial behavior (Allport, 1954). This research has linked guilt to a two-stage regulatory response: the initial experience of guilt is associated with dACC activation and a reduction in left prefrontal cortex (PFC)—a pattern associated with self-directed attention and behavioral inhibition, presumably to process one's misdeed and plan for reparation (Amodio et al., 2007; Fourie et al., 2014). This response then transforms into a state of readiness when an opportunity for reparation emerges, at which point one's initial feelings of guilt are associated with increased left PFC activity and the engagement of prejudice-reducing behaviors (Amodio et al., 2007). Several other emotions central to intergroup prejudice and behavior, such as disgust, hope, anger, pity, to name just a few, are ripe for further investigation.

Decision-making. Intergroup attitudes and emotions interact with other processes (e.g., valuation, stereotypes, social goals) to inform our social choices: whom to learn from, how to allocate our resources, how much to punish, and what norms to follow in social settings. A rapidly growing area of research in intergroup decision-making has begun to leverage knowledge acquired in the cognitive neuroscience of non-social learning and decision-making (see Ruff & Fehr, 2014 for review) to better understand how group contexts moderate these processes.

Conformity. We have already reviewed evidence that people exhibit greater sensorimotor resonance with ingroup relative to outgroup members experiencing pain, but it is crucial to understand whether other behaviors that rely on 'matching' a target's experience are sensitive to target group membership. For example, even chimps yawn more after watching video clips of familiar relative to unfamiliar conspecifics yawning (Campbell & DeWaal, 2011). To the extent that imitation is a rudimentary form of learning, such results suggest that people learn more from

ingroup than outgroup members. Recent findings comport with this prediction. In one study, participants rated a series of images on their valence, from negative to positive (Lin et al., 2018). Then, during an fMRI scan, American participants observed ratings of those same images ostensibly from other American and Chinese participants. Participants not only shifted their evaluations to conform more with ingroup relative to outgroup members' ratings, but this conformity behavior correlated with increased mPFC, left amygdala, left VS, bilateral AI, and bilateral ventrolateral prefrontal cortex responses—regions associated with positive valuation and value integration. Based on these results, the authors argued that rather than reflecting mere signaling strategy, conformity with the ingroup (or, distinguishing oneself from the outgroup; Huang et al., 2019) carries intrinsic value.

Moral judgments and punishment. Not all victims and perpetrators are equivalent; our judgments of wrongdoing are often modulated by targets' group memberships. Although there is a wealth of literature examining the neural substrates of moral decision-making, this work has only recently integrated considerations of group membership. For example, participants in an fMRI study reported being more upset when the victim of physical harm was a fellow university student (relative to a student from a rival university), but only when the perpetrator of harm was an outgroup member (i.e., a student from the rival university; Molenberghs et al., 2014). Only one region was associated with this moral response—left orbitofrontal cortex—which the authors speculated may support increased moral sensitivity by upregulating AI and amygdala responses to this special class of scenarios.

And what of lesser transgressions? Violators of social norms are often (though not always) punished more severely if they are outgroup relative to ingroup members. Using transcranial magnetic stimulation, one study found it was possible to eliminate this group bias

among soccer fans by disrupting activity in right (but not left) temporo-parietal junction, a region associated with mentalizing. More specifically, they found that disrupting right temporo-parietal junction reduced retaliation intentions, suggesting a link between mentalizing and punishment motives (Baumgartner et al., 2013).

Resource allocation. Finally, harkening back to some of the early work on intergroup relations in social psychology, which examined effects of group membership on resource distribution (Tajfel & Turner, 1977), recent social neuroscience studies have begun to examine the neural systems that generate biased resource allocations between ingroup and outgroup members. In Krosch and Amodio's (2019) fMRI study, described above, the degree of anti-Black disparities in White participants' monetary allocations were associated with activity in a fusiform-striatum pathway; that is, smaller resource allocations to Black recipients were predicted by reduced activity in the fusiform face area while viewing those recipients, coupled with reduced activity in the striatum. The authors speculated, based on this pattern, that scarcity may induce a form of perceptual dehumanization of racial outgroup members, which then signals their devaluation during allocation decisions.

In the context of the refugee crisis, one study tested the relative effects of peer-driven norms of altruism and oxytocin administration on resource allocations to refugees (Marsh et al., 2017). Their results were moderated by participants' xenophobia: low xenophobia participants were more inclined to help refugees than natives, and oxytocin to these participants increased donations for both groups. High xenophobia participants, by contrast, gave more to refugees than natives only when oxytocin was combined with the activation of altruism norms. However, we would be remiss if we did not note the large related literature examining the role of oxytocin in ethnocentrism (De Dreu et al., 2011), in-group defense (De Dreu et al., 2010), and even outgroup

attack (Zhang et al., 2019), indicating oxytocin's nuanced and complex influence on intergroup processes. Findings such as these begin to describe the neural processes associated with intergroup resource allocation decisions and, by doing so, shed new light on the psychological processes involved.

Summary: Intergroup Perception, Emotion, and Decision Making. Social neuroscience research has refined our understanding of how prejudice influences the visual processing of faces, intergroup emotion, and decision-making processes, particularly as each type of response pertains to behavior. These findings set the stage for important work to come on how these processes drive the impact of prejudice on critical everyday outcomes such as hiring, housing, voting, medical recommendations and care, and conflict resolution.

Self-regulation of prejudice

Despite the ease with which prejudice forms and springs to mind, many people consciously object to prejudice and strive to respond in an egalitarian manner (Devine, 1989). This conflict—between biased impulses and egalitarian intentions—has long been recognized in social psychology (Allport, 1954), and interventions to enhance control are an effective short-term strategy for reducing prejudice (Burns et al., 2017). However, while behavioral research has identified many factors that promote control, it has not addressed some crucial questions about the prejudice control process: for example, how is control initiated? Does control involve more than one process? On which psychological and behavioral processes does control operate? And why are some people better at controlling prejudice than others? Our ability to develop effective interventions to reduce prejudice depends on answers to questions such as these.

Social neuroscience studies have shown that prejudice control involves multiple processes, and that a consideration of these processes provides a more comprehensive account of intergroup behavior (Figure 6). Early neuroscience research on the regulation of prejudice adapted a cognitive neuroscience model of control, whereby control comprises (a) a monitoring process, supported by dACC, which detects the activation of bias, and (b) a regulatory process, supported by lateral PFC, which implements an intended response (Botvinick et al., 2001). When the monitoring process registers conflict, it signals the regulatory system to initiate control. According to this model, prejudice control is initiated when a conflict is detected between an activated bias (e.g., a stereotype-driven response) and an intended alternative response (Amodio et al., 2004; Richeson et al., 2003). Moreover, this conflict monitoring process has been shown to operate without awareness, suggesting that prejudice control may be initiated rapidly, without conscious deliberation (Nieuwenhuis et al., 2001).

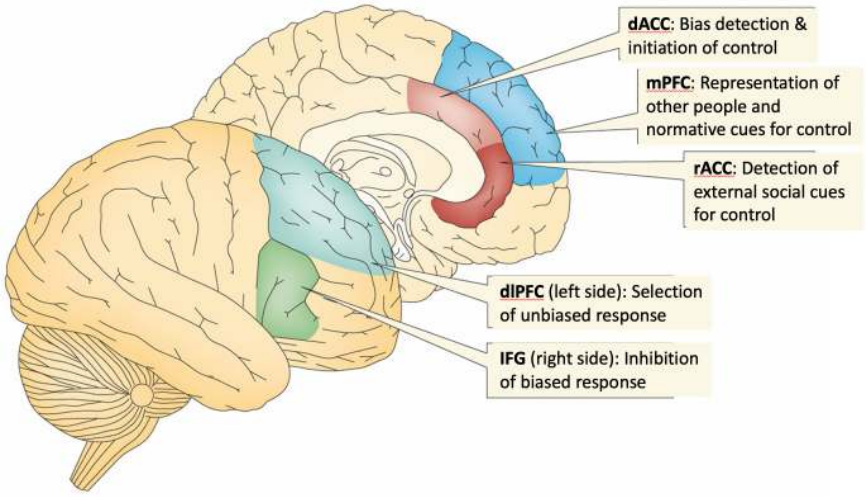

Figure 6. A model of prejudice control suggested by social neuroscience research, with descriptions of putative neural functions as they pertain specifically to the process of prejudice control. Medial regions (dACC, rACC, mPFC) support the detection of bias from both internal and external cues, whereas lateral regions (left dlPFC, right IFG) support the implementation of control via the selection or inhibition of responses. Adapted from Amodio (2014).

The conflict-detection component of prejudice control was tested in a study that assessed dACC activity in participants performing a task that required them to inhibit the automatic expression of racial stereotypes on some trials but not others (Amodio et al., 2004). Here, dACC was indexed by the error-related negativity (ERN) component of the ERP. ERN amplitudes were greater on trials requiring stereotype inhibition, and the magnitude of this neural signal predicted participants' success at controlling stereotype application in their behavior. Moreover, by demonstrating stereotype-related dACC activity on trials leading to both successful and unsuccessful control, this experiment dissociated the process of bias detection from the process of implementing a controlled response. Finally, by using an ERP index of dACC activity, which assesses changes in neural activity on the order of milliseconds, this work revealed that a neural signal to initiate control occurs rapidly (within about 300 ms) and thus likely without conscious deliberation.

This finding has been replicated and extended in several studies of prejudice control, using a variety of tasks and multiple ERP indices of dACC activity (Amodio et al., 2006; 2008; Amodio & Swencionis, 2018; Bartholow et al., 2006; Beer et al., 2008; Correll et al., 2006; Hughes et al., 2017). For instance, to address a prior finding that some people with egalitarian beliefs struggle to control automatic stereotypes more than others, one study showed that this individual difference in control could be explained by individuals' sensitivity to stereotype-based conflict, as indicated by dACC activity (Amodio et al., 2008). Other research has shown that personal and normative impetuses to control prejudice may rely on different mechanisms of conflict detection—a dACC process for detecting internal cues for control and an mPFC/rostral ACC process for monitoring external (e.g., social) cues—to explain why control based on external cues is often less effective than control based on internal cues (Amodio et al., 2006).

Hence, by distinguishing the conflict detection process as separate from the implementation of control, these studies provided novel accounts for enduring questions about prejudice control.

Similar effects have been observed using fMRI. In a study examining the neural correlates of the racial prejudice IAT—a task that requires controlled processing to complete bias-incompatible trials—dACC activity was associated with the ability to detect the correct, unbiased response amid biased automatic tendencies (Beer et al., 2008; see also Knutson et al., 2007). The role of dACC in the detection of potential bias was shown in an fMRI study by Norton et al. (2013), in which participants were asked to assign a stereotypic trait to one of a pair of target individuals. When targets in a pair differed in their race (one Black and one White), thereby creating the potential for stereotyping, participants slowed their response—a phenomenon the authors dubbed "racial paralysis." This reaction was associated with heightened dACC activity. dACC activity has even been observed during the passive viewing of racial outgroup faces, suggesting that the mere appearance of racial cues may engage a readiness for control (e.g., Cunningham et al., 2004; Richeson et al., 2003). Together, these studies demonstrate the involvement of the dACC in the detection of bias and the initiation of prejudice control, advancing our understanding of how control fails or succeeds.

Social neuroscience research has also shed new light on how control is implemented; that is, on what is being "controlled" during prejudice control. In several fMRI studies with White American participants, participants exhibited greater right IFG activity in response to presentations of Black faces compared with White faces (e.g., Beer et al., 2008; Cunningham et al., 2004; Mitchell et al., 2009; Richeson et al., 2003; Lieberman et al., 2005). Given research indicating that right IFG supports response inhibition (Aron et al., 2014), these findings suggest that exposure to Black faces elicited a form of behavioral inhibition. A similar pattern of right

IFG activity was observed when participants were asked to evaluate members of widely-stigmatized groups—a question that presumably requires the inhibition of a biased response (Krendl et al., 2009). Together, these findings suggest IFG supports an inhibitory form of prejudice control.

Whereas right IFG is associated with the inhibition of action, activity in the left lateral PFC has been associated with the production of goal-directed action. In the context of prejudice, this region has been linked to the successful implementation of an intended response over an automatic stereotype. In an EEG study designed to assess this process as it unfolded in real time (Amodio, 2010), brain activity was recorded in subjects as they completed stereotype priming task that, on some trials, required participants to replace an automatic stereotype response with a correct, unbiased response. Greater left dorsolateral prefrontal cortex (dlPFC) activity was associated with more success in overriding an automatic stereotype with an unbiased response. Furthermore, an analysis of ERPs during this process revealed that the effect of left dlPFC activity on stereotype control was mediated by rapid attentional orienting to racial outgroup cues, as indexed by the P2 component of the ERP. This pattern suggested that dlPFC activity tuned perceptual attention to relevant stimuli, in the manner of proactive control (e.g., Amodio & Swencionis, 2018), in order to promote the control of action. In another EEG study, noted above, greater left dlPFC activity was associated with participants' choice to engage in prejudice-reducing activities following a manipulation that made them feel guilty about their personal biases (Amodio et al., 2007).

These PFC findings suggest that, depending on the task, prejudice control may operate by inhibiting an unwanted behavioral response or by promoting goal-directed action to override an unwanted bias, or both, consistent with cognitive neuroscience models of PFC function (Miller

& Cohen, 2001). These expressions of control clarify and advance prior models of prejudice control that focused on correction, suppression, and inhibition (Amodio & Devine, 2010), or which assumed that control processes operated on internal mental representations rather than behavior. This model of control also updates an early view of prejudice control, whereby control was thought to operate via lateral PFC downregulation of the amygdala. Although this idea was suggested by some correlational findings, it is inconsistent with primate anatomical studies, which found sparse, if any, direct connections between these regions (Amodio & Ratner, 2011b; Gashghaei et al., 2007).

The new model of prejudice control suggested by social neuroscience has important implications for interventions to reduce prejudice. This model suggests that a prejudiced response may occur for multiple reasons, each associated with a different underlying process (Amodio, 2014). For example, a person may fail to detect the conflict between their biased response tendency and either their egalitarian goals or normative anti-prejudice social cues—processes that depends on the dACC or mPFC/rACC, respectively. Alternatively, they may have trouble inhibiting a biased response, despite having detected it—a process linked to right IFG. Or, they may have trouble identifying and implementing a desired egalitarian response—a processed supported by left dlPFC. As such, this model suggests that an intervention could target one or more of these specific processes. Moreover, different individuals may fail for different reasons and thus require different interventions. A consideration of these control processes and their relevance to subgroups of individuals promises a more refined and effective approach to prejudice reduction.

Summary: Self-Regulation of Prejudice. Considered together, social neuroscience research on prejudice control has significantly expanded psychological theory by identifying and

distinguishing multiple mechanisms of control (Figure 6). These include the detection of bias and initiation of control, in dACC—a process that can operate rapidly and in the absence of deliberation, and which can explain individual differences in prejudice control failures. This work has also elucidated mechanisms of control implementation, distinguishing between response inhibition, associated with right IFG, and the selection and application of intentional behavior, in left dlPFC. Together, these findings have advanced our understanding of the psychology of prejudice control and suggest new opportunities for prejudice reduction interventions.

Next questions and new challenges

When we consider the real-world effects of prejudice in society, it becomes obvious that social neuroscience research on prejudice still has much to do. To date, research from this field has focused on the psychological building blocks of prejudice—for example, processes of social categorization, prejudice formation, intergroup emotion and perception and, more recently, the neurocomputational basis of these processes. However, as this field continues to develop, it must make connections to the real life forms of prejudice that persist in society, from expressions of bias in real dyadic intergroup social interactions and the spread of prejudice across members of a community, to institutional discrimination, systematic forms of oppression such as voter suppression, and even ethnic conflict and genocide. These goals will require new methods, greater ecological validity, and increased collaboration with scientists and scholars from other disciplines.

Ambulatory (i.e., "wearable") neuroimaging technologies now make it possible to record participants' neural and physiological activity during direct social interaction, potentially increasing ecological validity and permitting real dyadic analysis. For instance, methods such as ambulatory EEG and functional near-infrared spectroscopy (fNIRs), in which participants wear a sensor cap but can otherwise move naturally, offer the possibility of examining neural activity during more naturalistic intergroup interaction. Furthermore, the enhanced study of dyadic interactions will elucidate the effects of an actor's prejudice on a target's response to being stigmatized (e.g., Welborn et al., in press). As these technologies develop, they will increasingly inform questions about the neural basis of real-world prejudice.

Questions about how information spreads across a social group and influences its members' behaviors have recently been examined using network analysis (Weaverdyck & Parkinson, 2018). Such methods examine similarities in patterns of brain activity across members of a group and compare them with patterns of judgments toward other group members. Similar methods can address the spread of prejudice and stereotypes within a community, potentially informing the connection between individual-level neural activations and group-level processes (Parkinson & Du, 2020).

Finally, researchers have begun to examine the neural processes involved in real-world intergroup conflict. One fMRI study examined neural activity of White and Black South Africans viewing testimony from the Truth and Reconciliation Commission on their experiences under Apartheid—an extremely emotional event that elicited outwardly egalitarian behaviors among White participants despite their pro-ingroup patterns of neural activity (Fourie et al., 2017). Other research has begun to examine the neural roots of dehumanization as it relates to real-world national and ethnic conflict (Bruneau et al., 2018). The broader goal of this work is to

identify ways to apply knowledge of the neuroscience of prejudice to interventions to reduce intergroup animus and conflict. Hence, in our view, the most critical questions and challenges facing this field in the next decade concern its ability to connect basic neurocognitive process to a broader array of intergroup contexts, factors, and outcomes.

Conclusions

The social neuroscience of prejudice is a rich and thriving area of research that addresses questions about the psychology of prejudice with the tools of cognitive neuroscience and psychophysiology. Here, we have highlighted major theoretical advances produced by this literature to date: from the way in which we perceive groups, to how prejudice is learned and represented in the mind, how it influences our perceptions, emotions, and decisions, and how it can be regulated. By applying theories and tools of neuroscience to this complex social phenomenon, this work has produced a more refined understanding of the psychological processes involved in prejudice, along with new insights and theoretical connections that might not have emerged from traditional behavioral approaches. Nevertheless, this area of research is still relatively new; as the fields of cognitive neuroscience and intergroup psychology continue to evolve and advance, so too will the social neuroscience of prejudice. In this field marked by innovation, we look forward to the new discoveries that will further our understanding of prejudice and its role in social behavior.

References

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

Amodio, D. M. (2008). The social neuroscience of intergroup relations. *European Review of Social Psychology*, *19*, 1-54.

Amodio, D. M. (2009). Intergroup anxiety effects on the control of racial stereotypes: A psychoneuroendocrine analysis. *Journal of Experimental Social Psychology*, *45*, 60-67.

Amodio, D. M. (2010). Coordinated roles of motivation and perception in the regulation of intergroup responses: Frontal cortical asymmetry effects on the P2 event-related potential and behavior. *Journal of Cognitive Neuroscience*, *22*, 2609-2617.

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*(10), 670.

Amodio, D. M. (2019). Social Cognition 2.0: An interactive memory systems account. *Trends in Cognitive Sciences*, 23, 21-33.

Amodio, D. M., Bartholow, B. D., & Ito, T. A. (2014). Tracking the dynamics of the social brain: ERP approaches for social cognitive & affective neuroscience. *Social Cognitive & Affective Neuroscience, 9,* 385-393.

Amodio, D. M. & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652-661.

Amodio, D. M. & Devine, P. G. (2010). Regulating behavior in the social world: Control in the context of intergroup bias. In R. R. Hassin, K. N. Ochsner, and Y. Trope (Eds). *Self control in society, mind and brain* (pp. 49-75). New York: Oxford University Press.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science*, *18*, 524-530.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94,* 60-74.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*, 268-277.

Amodio, D. M., & Hamilton, H. K. (2012). Intergroup anxiety effects on implicit racial evaluation and stereotyping. *Emotion, 12,* 1273-1280.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink responses and self-report. *Journal of Personality and Social Psychology, 84*, 738-753.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15,* 88-93.

Amodio, D. M., Kubota, J. T., Harmon-Jones, E., & Devine, P. G. (2006). Alternative mechanisms for regulating racial responses according to internal vs. external cues. *Social Cognitive and Affective Neuroscience*, *1*, 26-36.

Amodio, D. M., & Ratner, K. G. (2011a). A memory systems model of implicit social cognition. *Current Directions in Psychological Science, 20,* 143-148.

Amodio, D. M., & Ratner, K. (2011b). Mechanisms for the regulation of intergroup responses: A social neuroscience analysis. In J. Decety and J. T. Cacioppo (Eds.), *Handbook of social neuroscience* (pp. 729-741). New York: Oxford University Press.

Amodio, D. M., & Swencionis, J. K. (2018). Proactive control of implicit bias: A theoretical model and implications for behavior change. *Journal of Personality and Social Psychology, 115,* 255-275.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: one decade on. *Trends in cognitive sciences*, *18*(4), 177-185.

Avenanti, A., Sirigu, A., & Aglioti, S. M. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, *20*, 1018-1022.

Azevedo, R. T., Macaluso, E., Avenanti, A., Santangelo, V., Cazzato, V., & Aglioti, S. M. (2013). Their pain is not our pain: brain and autonomic correlates of empathic resonance with the pain of same and different race individuals. *Human brain mapping*, *34*, 3168-3181.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological bulletin*, *140*, 1556-1581.

Bartholow, B. D., Dickter, C. L., & Sestir, M. A. (2006). Stereotype activation and control of race bias: Cognitive control of inhibition and its impairment by alcohol. *Journal of personality and social psychology*, *90*, 272-287.

Bartholow, B. D., Fabiani, M., Gratton, G., & Bettencourt, B. A. (2001). A psychophysiological examination of cognitive processing of and affective responses to social expectancy violations. *Psychological science*, *12*, 197-204.

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R., & Knoch, D. (2013). Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Social cognitive and affective neuroscience*, *9*, 653-660.

Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *Neuroimage*, *43*, 775-783.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *science*, *324*, 1160-1164.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624-652.

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of personality and social psychology*, *47*, 1191-1205.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of social issues*, *55*(3), 429-444.

Brosch, T., Bar-David, E., & Phelps, E. A. (2013). Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychological science*, *24*, 160-166.

Brown, L. M., Bradley, M. M., & Lang, P. J. (2006). Affective reactions to pictures of ingroup and outgroup members. *Biological psychology*, *71*, 303-311.

Bruneau, E. G., Dufour, N., & Saxe, R. (2012). Social cognition in members of conflict groups: behavioural and neural responses in Arabs, Israelis and South Americans to each other's misfortunes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 717-730.

Bruneau, E., Jacoby, N., Kteily, N., & Saxe, R. (2018). Denying humanity: The distinct neural correlates of blatant dehumanization. *Journal of Experimental Psychology: General, 147*(7), 1078–1093. https://doi.org/10.1037/xge0000417

Burns, M. D., Monteith, M. J., & Parker, L. R. (2017). Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application. *Journal of Experimental Social Psychology*, *73*, 97-110.

Caldara, R., Thut, G., Servoir, P., Michel, C. M., Bovet, P., & Renault, B. (2003). Face versus non-face object perception and the 'other-race'effect: a spatio-temporal event-related potential study. *Clinical Neurophysiology*, *114*, 515-528.

Campbell, M. W., & De Waal, F. B. (2011). Ingroup-outgroup bias in contagious yawning by chimpanzees supports link to empathy. *PloS one*, *6*(4), e18283.

Delgado, M. R., Olsson, A., & Phelps, E. A. (2006). Extending animal models of fear conditioning to humans. *Biological psychology*, *73*, 39-48.

Cassidy, K. D., Boutsen, L., Humphreys, G. W., & Quinn, K. A. (2014). Ingroup categorization affects the structural encoding of other-race faces: Evidence from the N170 event-related potential. *Social neuroscience*, *9*, 235-248.

Chekroud, A. M., Everett, J. A., Bridge, H., & Hewstone, M. (2014). A review of neuroimaging studies of race-related prejudice: does amygdala response reflect threat? *Frontiers in Human Neuroscience*, *8*, 179.

Cikara, M. (2015). Intergroup schadenfreude: Motivating participation in collective violence. *Current opinion in behavioral sciences*, *3*, 12-17.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science*, *9*(3), 245-274.

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*(3), 306-313.

Cikara, M., Bruneau, E. G., & Saxe, R. R. (2011). Us and them: Intergroup failures of empathy. *Current Directions in Psychological Science*, *20*(3), 149-153.

Cikara, M., Van Bavel, J. J., Ingbretsen, Z. A., & Lau, T. (2017). Decoding "us" and "them": Neural representations of generalized group concepts. *Journal of Experimental Psychology: General*, *146*(5), 621.

Clark, V.P., Fan, S., & Hillyard, S.A. (1995). Identification of early visual evoked potential generators by retinotopic and topographic analyses. *Human Brain Mapping, 2*, 170-187.

Contreras-Huerta, L. S., Baker, K. S., Reynolds, K. J., Batalha, L., & Cunnington, R. (2013). Racial bias in neural empathic responses to pain. *PloS one*, *8*(12), e84001.

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social cognitive and affective neuroscience*, *7*, 764-770.

Correll, J., Urland, G. R., & Ito, T. A. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology*, *42*, 120-128.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological science*, *15*, 806-813.

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., ... & Feith, S. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science*, *328*(5984), 1408-1411.

De Dreu, C. K., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences*, *108*(4), 1262-1266.

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

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.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, *56*(1), 5-18.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of personality and social psychology*, *82*(5), 835-848.

Dickter, C. L., & Bartholow, B. D. (2007). Racial ingroup and outgroup attention biases revealed by event-related brain potentials. *Social cognitive and affective neuroscience*, *2*, 189-198.

Dovidio JF, Gaertner SL. 2010. Intergroup bias. In *Handbook of Social Psychology*, ed. ST Fiske, DT Gilbert, G Lindzey, pp. 1084–121. Hoboken, NJ: Wiley

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology*, *82*(1), 62-68.

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual Review of Vision Science*, *1*, 393-416.

Dunsmoor, J. E., Kubota, J. T., Li, J., Coelho, C. A., & Phelps, E. A. (2016). Racial stereotypes impair flexibility of emotional learning. *Social cognitive and affective neuroscience*, *11*, 1363-1373.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?. *Journal of personality and social psychology*, *69*, 1013-1027.

Feng, L., Liu, J., Wang, Z., Li, J., Li, L., Ge, L., ... & Lee, K. (2011). The other face of the other-race effect: An fMRI investigation of the other-race face categorization advantage. Neuropsychologia, 49(13), 3739-3749.

Filion, D. L., Dawson, M. E., & Schell, A. M. (1998). The psychological significance of human startle eyeblink modification: a review. *Biological psychology*, *47*(1), 1-43.

Fincher, K. M., & Tetlock, P. E. (2016). Perceptual dehumanization of faces is activated by norm violations and facilitates norm enforcement. *Journal of Experimental Psychology: General*, *145*(2), 131-146.

Fini, C., Cardini, F., Tajadura-Jiménez, A., Serino, A., & Tsakiris, M. (2013). Embodying an outgroup: the role of racial bias and the effect of multisensory processing in somatosensory remapping. *Frontiers in behavioral neuroscience*, *7*, 165.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, *2*(4), 357-411.

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

Foerde, K. (2018). What are habits and do they depend on the striatum? A view from the study of neuropsychological populations. *Current opinion in behavioral sciences*, *20*, 17-24.

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, *45*, 152-170.

Fourie, M. M., Stein, D. J., Solms, M., Gobodo-Madikizela, P., & Decety, J. (2017). Empathy and moral emotions in post-apartheid South Africa: an fMRI investigation. *Social cognitive and affective neuroscience*, *12*, 881-892.

Fourie, M. M., Thomas, K. G. F., Amodio, D. M., Warton, C. M. R., & Meintjes, E. M. (2014). Neural correlates of experienced moral emotion: An fMRI investigation of emotion in response to prejudice feedback. *Social Neuroscience*, *9*, 203-218.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in cognitive sciences*, *20*, 362-374.

Freeman, J.B., Ambady, N., & Holcomb, P.J. (2010). The face-sensitive N170 encodes social category information. *NeuroReport, 21*, 24-28.

Gallate, J., Wong, C., Ellwood, S., Chi, R., & Snyder, A. (2011). Noninvasive brain stimulation reduces prejudice scores on an implicit association test. *Neuropsychology*, *25*(2), 185-192.

Gershman, S. J., & Cikara, M. (in press). Social structure learning. *Current Directions in Psychological Science*.

Ghashghaei, H. T., Hilgetag, C. C., & Barbas, H. (2007). Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala. *Neuroimage*, *34*, 905-923.

Gilbert, S. J., Swencionis, J. K., & Amodio, D. M. (2012). Evaluative vs. trait representation in intergroup social judgments: Distinct roles of anterior temporal lobe and prefrontal cortex. *Neuropsychologia, 50*, 3600-3611.

Golby, A. J., Gabrieli, J. D., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature neuroscience*, *4*(8), 845-850.

Gutsell, J. N., & Inzlicht, M. (2010). Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups. *Journal of experimental social psychology*, *46*, 841-845.

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology, 24,* 92-97.

Hackel, L. M., Berg, J. J., Lindström, B. R., & Amodio, D. M. (2019). Model-Based and Model-Free Social Cognition: Investigating the role of habit in social attitude formation and choice. *Frontiers in Psychology, 10, article 2592*. doi: 10.3389/fpsyg.2019.02592

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on decision making. *Nature Neuroscience, 18,* 1233-1235.

Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*.

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: Volatile ambivalence. *Envy: Theory and research*, 133-147.

He, Y., Johnson, M. K., Dovidio, J. F., & McCarthy, G. (2009). The relation between race-related implicit associations and scalp-recorded neural activity evoked by faces from different races. *Social Neuroscience*, *4*, 426-442.

Hein, G., Engelmann, J. B., Vollberg, M. C., & Tobler, P. N. (2016). How learning shapes the empathic brain. *Proceedings of the National Academy of Sciences*, *113*, 80-85.

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.

Hein, G., & Singer, T. (2008). I feel how you feel but not always: the empathic brain and its modulation. *Current opinion in neurobiology*, *18*(2), 153-158.

Holland, P. C., & Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends in cognitive sciences*, *3*, 65-73.

Huang, Y., Zhen, S., & Yu, R. (2019). Distinct neural patterns underlying ingroup and outgroup conformity. *Proceedings of the National Academy of Sciences*, *116*(11), 4758-4759.

Hughes, B. L., Ambady, N., & Zaki, J. (2017). Trusting outgroup, but not ingroup members, requires control: neural and behavioral evidence. *Social Cognitive and Affective Neuroscience*, *12*, 372-381.

Hughes, B. L., Camp, N. P., Gomez, J., Natu, V. S., Grill-Spector, K., & Eberhardt, J. L. (2019). Neural adaptation to faces reveals racial outgroup homogeneity effects in early perception. *Proceedings of the National Academy of Sciences*, *116*(29), 14532-14537.

Ibáñez, A., Gleichgerrcht, E., Hurtado, E., González, R., Haye, A., & Manes, F. F. (2010). Early neural markers of implicit attitudes: N170 modulated by intergroup and evaluative contexts in IAT. *Frontiers in Human Neuroscience*, *4*, 188.

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

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of personality and social psychology*, *85*(4), 616.

Ito, T. A., & Urland, G. R. (2005). The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(1), 21-36.

Ito, T. A., & Tomelleri, S. (2017). Seeing is not stereotyping: the functional independence of categorization and stereotype activation. *Social cognitive and affective neuroscience*, *12*, 758-764.

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

Kawakami, K., Amodio, D. M., & Hugenberg, K. (2017). Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. *Advances in Experimental Social Psychology, 55*, 1-80.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399-1402.

Knutson, K. M., Mah, L., Manly, C. F., & Grafman, J. (2007). Neural correlates of automatic beliefs about gender and race. *Human brain mapping*, *28*, 915-930.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/ socioeconomic success, conservative–progressive beliefs, and communion. Journal of Personality and Social Psychology, 110, 675–709.

Krendl, A. C., Heatherton, T. F., & Kensinger, E. A. (2009). Aging minds and twisting attitudes: An fMRI investigation of age differences in inhibiting prejudice. *Psychology and aging*, *24*, 530-541.

Krosch, A. R., & Amodio, D. M. (2019). Scarcity disrupts the neural encoding of Black faces: A socioperceptual pathway to discrimination. *Journal of personality and social psychology*, *117*, 859-875.

Kubota, J. T., & Ito, T. A. (2007). Multiple cues in social perception: The time course of processing race and facial expression. *Journal of Experimental Social Psychology*, *43*, 738-752.

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

Kurdi, B., Gershman, S. J., & Banaji, M. R. (2019). Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *Proceedings of the National Academy of Sciences*, *116*, 6035-6044.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621-647.

Lamm, C., Rütgen, M., & Wagner, I. C. (2019). Imaging empathy and prosocial emotions. *Neuroscience letters*, *693*, 49-53.

Lau, T., Gershman, S. J., & Cikara, M. (2020). Social structure learning in human anterior insula. *eLife*, *9*, e53162.

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, *147*(12), 1881.

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, *73*(4), 653-676.

LeDoux, J. E., & Hofmann, S. G. (2018). The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences*, *19*, 67-72.

Levy, J., Goldstein, A., Influs, M., Masalha, S., Zagoory-Sharon, O., & Feldman, R. (2016). Adolescents growing up amidst intractable conflict attenuate brain response to pain of outgroup. *Proceedings of the National Academy of Sciences*, *113*(48), 13696-13701.

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature neuroscience*, *8*, 720-722.

Lin, L. C., Qu, Y., & Telzer, E. H. (2018). Intergroup social influence on emotion processing in the brain. *Proceedings of the National Academy of Sciences*, *115*(42), 10630-10635.

Mackie, D. M., & Smith, E. R. (2018). Intergroup emotions theory: Production, regulation, and modification of group-based emotions. In *Advances in experimental social psychology* (Vol. 58, pp. 1-69). Academic Press.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual review of psychology*, *51*, 93-120.

Mallan, K. M., Sax, J., & Lipp, O. V. (2009). Verbal instruction abolishes fear conditioned to racial out-group faces. *Journal of Experimental Social Psychology*, *45*, 1303–1307.

March, D. S., Gaertner, L., & Olson, M. A. (2018). On the prioritized processing of threat in a dual implicit process model of evaluation. *Psychological Inquiry*, *29*, 1-13.

Marsh, N., Scheele, D., Feinstein, J. S., Gerhardt, H., Strang, S., Maier, W., & Hurlemann, R. (2017). Oxytocin-enforced norm compliance reduces xenophobic outgroup rejection. *Proceedings of the National Academy of Sciences*, *114*(35), 9314-9319.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25-45.

Mathur, V. A., Harada, T., Lipke, T., & Chiao, J. Y. (2010). Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage*, *51*(4), 1468-1475.

Mattan, B. D., Kubota, J. T., Dang, T. P., & Cloutier, J. (2018). External motivation to avoid prejudice alters neural responses to targets varying in race and status. *Social cognitive and affective neuroscience*, *13*, 22-31.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of experimental Social psychology*, *37*, 435-442.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, *214*, 655-667.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*, 167-202.

Mitchell, J. P., Ames, D. L., Jenkins, A. C., & Banaji, M. R. (2009). Neural correlates of stereotype application. *Journal of cognitive neuroscience*, *21*, 594-604.

Molapour, T., Golkar, A., Navarrete, C. D., Haaker, J., & Olsson, A. (2015). Neural correlates of biased social fear learning and interaction in an intergroup context. *NeuroImage*, *121*, 171-183.

Molenberghs, P., & Morrison, S. (2012). The role of the medial prefrontal cortex in social categorization. *Social cognitive and affective neuroscience*, *9*(3), 292-296.

Molenberghs, P., Gapp, J., Wang, B., Louis, W. R., & Decety, J. (2014). Increased moral sensitivity for outgroup perpetrators harming ingroup members. *Cerebral Cortex*, *26*(1), 225-233.

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

Navarrete, C. D., McDonald, M. M., Asher, B. D., Kerr, N. L., Yokota, K., Olsson, A., & Sidanius, J. (2012). Fear is readily associated with an out-group face in a minimal group context. *Evolution and Human Behavior*, *33*, 590–593.

Navarrete, C. D., Olsson, A., Ho, A. K., Mendes, W. B., Thomsen, L., & Sidanius, J. (2009). Fear extinction to an out-group face: The role of target gender. *Psychological Science*, *20*, 155–158. doi:10.1111/j.1467-9280.2009.02273

Neuberg, S. L., & Schaller, M. (2016). An evolutionary threat-management approach to prejudices. *Current Opinion in Psychology*, *7*, 1-5.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus—norepinephrine system. *Psychological bulletin*, *131*, 510-532.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, *38*, 752-760.

Norton, M. I., Mason, M. F., Vandello, J. A., Biga, A., & Dyer, R. (2013). An fMRI investigation of racial paralysis. *Social cognitive and affective neuroscience*, *8*, 387-393.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual review of psychology*, *68*, 73-100.

Ofan, R. H., Rubin, N., Amodio, D. M. (2011). Seeing race: N170 responses to race and their relation to automatic racial attitudes and controlled processing. *Journal of Cognitive Neuroscience*, *23*, 3152-3161.

Ofan, R. H., Rubin, N., Amodio, D. M. (2014). Situation-based social anxiety enhances the neural encoding of faces: Evidence from an intergroup context. *Social Cognitive & Affective Neuroscience, 9,* 1055-1061.

Öhman, A., & Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of"preparedness"? Journal of Personality and Social Psychology, 36, 1251-1258.

Olson, I. R., McCoy, D., Klobusicky, E., & Ross, L. A. (2013). Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social cognitive and affective neuroscience*, *8*, 123-133.

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science, 309*, 785–787.

Parkinson, C., & Du, M. (2020). How does the brain infer hidden structures? *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2020.05.002

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of cognitive neuroscience*, *12*, 729-738.

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.

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience & Biobehavioral Reviews*, *32*, 197-205.

Quadflieg, S., Turk, D. J., Waiter, G. D., Mitchell, J. P., Jenkins, A. C., & Macrae, C. N. (2009). Exploring the neural correlates of social stereotyping. *Journal of Cognitive Neuroscience*, *21*, 1560-1570.

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences*, *114*, 8511-8516.

Ratner, K. G., & Amodio, D. M. (2013). Seeing “us vs. them": Minimal group effects on the neural encoding of faces. *Journal of Experimental Social Psychology, 49, 298-301.*

Ratner, K. G., Kaul, C., & Van Bavel, J. J. (2013). Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive and Affective Neuroscience*, *8*, 750-755.

Reggev, N., Brodie, K., Cikara, M., & Mitchell, J. P. (2020). Human face-selective cortex does not distinguish between members of a racial out-group**.** *eNeuro*.

Richeson, J. A., & Trawalter, S. (2008). The threat of appearing prejudiced and race-based attentional biases. *Psychological Science*, *19*, 98-102.

Richeson, J. A., Baird, A. A., Gordon, H. L., Heatherton, T. F., Wyland, C. L., Trawalter, S., & Shelton, J. N. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature neuroscience*, *6*, 1323-1328.

Richeson, J. A., Todd, A. R., Trawalter, S., & Baird, A. A. (2008). Eye-gaze direction modulates race-related amygdala activity. *Group Processes & Intergroup Relations*, *11*, 233-246.

Riečanský, I., Paul, N., Kölble, S., Stieger, S., & Lamm, C. (2014). Beta oscillations reveal ethnicity ingroup bias in sensorimotor resonance to pain of others. *Social cognitive and affective neuroscience*, *10*(7), 893-901.

Robbins, T. W., & Costa, R. M. (2017). Habits. *Current biology*, *27*, R1200-R1206.

Ronquillo, J., Denson, T. F., Lickel, B., Lu, Z. L., Nandy, A., & Maddox, K. B. (2007). The effects of skin tone on race-related amygdala activity: An fMRI investigation. *Social cognitive and affective neuroscience*, *2*(1), 39-44.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport*, *11*, 69-72.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549.

Saxe, R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. *Handbook of theory of mind*, 1-35.

Scheepers, D., Derks, B., Nieuwenhuis, S., Lelieveld, G. J., Van Nunspeet, F., Rombouts, S. A., & De Rover, M. (2013). The neural correlates of in-group and self-face perception: is there overlap for high identifiers?. *Frontiers in human neuroscience*, *7*, 528.

Schmid, P. C., & Amodio, D. M. (2017). Power effects on implicit prejudice and stereotyping: The role of intergroup face processing. *Social Neuroscience, 12*, 218-231.

Senholzi, K. B., & Ito, T. A. (2013). Structural face encoding: how task affects the N170's sensitivity to race. *Social Cognitive and Affective Neuroscience*, *8*, 937-942.

Sheng, F., Liu, Q., Li, H., Fang, F., & Han, S. (2014). Task modulations of racial bias in neural responses to others' suffering. *NeuroImage*, *88*, 263-270.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Sidanius, J., & Pratto, F. (2012). Social dominance theory. In Lange, Paul A.M., Kruglanski, Arie W. and Higgins. E. T. (Eds.) In Handbook of Theories of Social Psychological. Pp. 418-439. London: Sage Publications.

Spiers, H. J., Love, B. C., Le Pelley, M. E., Gibb, C. E., & Murphy, R. A. (2017). Anterior temporal lobe tracks the formation of prejudice. *Journal of cognitive neuroscience*, *29*(3), 530-544.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the United States of America, 93*, 13515–13522.

Stahl, J., Wiese, H., & Schweinberger, S. R. (2008). Expertise and own-race bias in face processing: an event-related potential study. *Neuroreport*, *19*, 583-587.

Stanley, D. A., Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., & Phelps, E. A. (2012). Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 744-753.

Stolier, R.M. & Freeman, J.B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. Nature Neuroscience, 19, 795-797.

Stolier, R.M. & Freeman, J.B. (2017). A neural mechanism of social categorization. Journal of Neuroscience, 37, 5711-5721.

Telzer, E. H., Humphreys, K. L., Shapiro, M., & Tottenham, N. (2013). Amygdala sensitivity to race is not present in childhood but emerges over adolescence. *Journal of cognitive neuroscience*, *25*, 234-244.

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

Trawalter, S., Adam, E. K., Chase-Lansdale, P. L., & Richeson, J. A. (2012). Concerns about appearing prejudiced get under the skin: Stress responses to interracial contact in the moment and across time. *Journal of Experimental Social Psychology*, *48*, 682-693.

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.

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*, 1131-1139.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience*, *23*, 3343-3354.

Vanman, E. J., Ryan, J. P., Pedersen, W. C., & Ito, T. A. (2013). Probing prejudice with startle eyeblink modification: A marker of attention, emotion, or both. *International journal of psychological research*, *6*, 30-41.

Vizioli, L., Foreman, K., Rousselet, G. A., & Caldara, R. (2010). Inverting faces elicits sensitivity to race on the N170 component: A cross-cultural study. *Journal of Vision, 10,* 1-23.

Vollberg, M. C., & Cikara, M. (2018). The neuroscience of intergroup emotion. *Current opinion in psychology*, *24*, 48-52.

Volpert-Esmond, H. I., & Bartholow, B. D. (2019). Explicit categorization goals affect attention-related processing of race and gender during person construal. *Journal of Experimental Social Psychology*, *85*, 103839.

Walker, P. M., Silvert, L., Hewstone, M., & Nobre, A. C. (2008). Social contact and other-race face processing in the human brain. *Social Cognitive and Affective Neuroscience*, *3*, 16-25.

Weaverdyck, M. E., & Parkinson, C. (2018). The neural representation of social networks. *Current opinion in psychology*, *24*, 58-66.

Welborn, B. L., Hong, Y., & Ratner, K. G. (2020). Exposure to negative stereotypes influences representations of monetary incentives in the nucleus accumbens. *Social Cognitive and Affective Neuroscience*.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, *16*, 56-63.

White, K. R., Crites Jr, S. L., Taylor, J. H., & Corral, G. (2009). Wait, what? Assessing stereotype incongruities using the N400 ERP component. *Social Cognitive and Affective Neuroscience*, *4*, 191-198.

Wiese, H., Stahl, J., & Schweinberger, S. R. (2009). Configural processing of other-race faces is delayed but not decreased. *Biological psychology*, *81*, 103-109.

Willadsen-Jensen, E. C., & Ito, T. A. (2008). A foot in both worlds: Asian Americans' perceptions of Asian, White, and racially ambiguous faces. *Group Processes & Intergroup Relations*, *11*(2), 182-200.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological review*, *114*, 843-863.

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience*, *29*(26), 8525-8529.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences*, *104*(15), 6430-6435.

Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. *Nature neuroscience*, *15*(5), 675.

Zhang, H., Gross, J., De Dreu, C., & Ma, Y. (2019). Oxytocin promotes coordinated out-group attack during intergroup conflict in humans. *Elife*, *8*, e40698.

Zheng, X., & Segalowitz, S. J. (2014). Putting a face in its place: In-and out-group membership alters the N170 response. *Social cognitive and affective neuroscience*, *9*, 961-968.

Zhou, Y., Gao, T., Zhang, T., Li, W., Wu, T., Han, X., & Han, S. (2020). Neural dynamics of racial categorization predicts racial bias in face recognition and altruism. *Nature human behaviour*, *4*(1), 69-87.

Glossary

Prejudice: negative evaluation of a social group and its generalization to group members

Social cognition: Field of psychology concerned with the cognitive processes through which we perceive, think about, judge, and act toward people and in response to social contexts.

Social Neuroscience: field of research that probes the connection between the brain and social behavior

Event-related potential (ERP): pattern of neural activity, measured using electroencephalography, that are time-locked to a specific event (e.g., stimulus or response).

Multi-voxel pattern analysis (MVPA): unlike traditional univariate analysis, MVPA allows investigators to examine patterns of voxel-level neural activation within a brain region (which may go undetected by traditional univariate analysis) and use them to differentiate cognitive representations.

Transcranial magnetic stimulation (TMS): non-invasive form of brain stimulation; uses electromagnetic fields to manipulate neural activity in a circumscribed brain region.

Acronyms

ATL: anterior temporal lobe

dACC: dorsal anterior cingulate cortex

dlPFC: dorsolateral prefrontal cortex

ERN: error-related negativity ERP component

FFA: fusiform face area

IFG: inferior frontal gyrus

mPFC: medial prefrontal cortex

MTL: medial temporal lobe

VS: ventral striatum

vmPFC: ventral medial prefrontal cortex