The Neural Underpinnings of Intergroup Bias

David M. Amodio[1, 2]

Jeffrey J. Berg[1]

[1]New York University, [2]University of Amsterdam

Amodio, D. M. & Berg, J. J. (in press). The Neural Underpinnings of Intergroup Bias. To appear in P. Molenberghs (Ed.) *The Neuroscience of Intergroup Relations*. New York: Routledge.

Correspondence to:

David M. Amodio
Department of Psychology
New York University
6 Washington Place
New York, NY 10003
david.amodio@nyu.edu

Abstract

Intergroup bias is a pervasive feature of human life, borne of our reliance on group living, which represents the interplay of social structures, group dynamics, and the minds and behaviors of individuals. This complexity is reflected in its neural basis: the social neuroscience of intergroup bias examines how multiple neural systems operate in concert to support high-level social responses and the ability to coordinate with groups and societies. This chapter describes what we have learned so far about the neural basis of how intergroup bias is represented in the mind, integrated into a memory systems model of intergroup bias. We discuss the implications of this model for how intergroup bias is expressed in behavior and how it may be reduced, as well as how this approach may begin to illuminate the interface between systemic elements of bias individual minds.

How can neuroscience—the study of neural processes—help us to understand the complex social phenomenon of intergroup bias? Historically, group-based biases have been understood in terms of systemic- and individual-level factors. At the system level, intergroup bias is instantiated in social hierarchies and power structures that promote and maintain group disparities. At the individual level, intergroup bias may manifest in a person's stereotypic beliefs, negative attitudes and affective responses, and discriminative actions, and may be expressed collectively in large-scale patterns of discrimination and campaigns of oppression. Moreover, these two aspects of intergroup bias—systemic and individual—can interact in a cyclical manner, whereby systemic disparities are mediated by individual biases, which in turn guide discriminatory actions that reinforce existing disparities.

Although intergroup bias is typically understood in terms of these psychological and sociological processes, research on the neural underpinnings of intergroup bias has begun to shed light on their underlying cognitive and affective processes, offering insight into the specific ways that intergroup bias is experienced and expressed among individuals (Amodio, 2014; Molenberghs, 2013). Recent advances in social neuroscience have also begun to address how individual minds interact with broader elements of their social system, suggesting that social neuroscience can also offer insight into broader psychosocial mechanisms of intergroup bias.

In this chapter, we present a social neuroscience model of the processes through which intergroup bias is learned, mentally represented, and expressed in behavior. We begin by describing the *memory systems model* of intergroup bias (Amodio & Ratner, 2011; see also Amodio, 2019), which applies research on learning and memory established in cognitive neuroscience to the social phenomenon of intergroup bias, and discuss its implications for how

intergroup bias operates in the individual level, as well as how individual-level biases may interface with social systems.

A memory systems model of intergroup bias

Psychological models of intergroup bias and social cognition have traditionally assumed that we learn about people and groups in one particular way: through the formation of conceptual knowledge associations that are represented in a single, broad semantic network. This assumption underlies influential dual-process models of prejudice and stereotyping, and it continues to guide much thinking on how intergroup bias is formed, expressed, and potentially reduced via intervention (see Amodio, 2019, for a review). However, this view is quite different from models developed in cognitive neuroscience, which identify multiple forms of learning and memory. These memory systems are distinguished by their neural substrates, the kinds of information encoded (e.g., conceptual or reward-based), their mode of updating, and their expression in judgment and behavior (Poldrack & Foerde, 2008; Squire & Zola, 1996).

We have argued that a consideration of memory systems is crucial to our understanding of intergroup bias because it suggests that multiple kinds of information are encoded, beyond semantic knowledge, and that these different kinds of information are expressed in different channels of intergroup social behavior (Amodio, 2019; Amodio & Berg, 2018; Amodio & Ratner, 2011). These systems include memory processes addressed in traditional stereotyping and 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. A sample of these learning and memory systems is shown in Figure 1, 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 neuroscientific models of learning and memory, and discuss their implications for how biases may be activated and expressed in behavior.

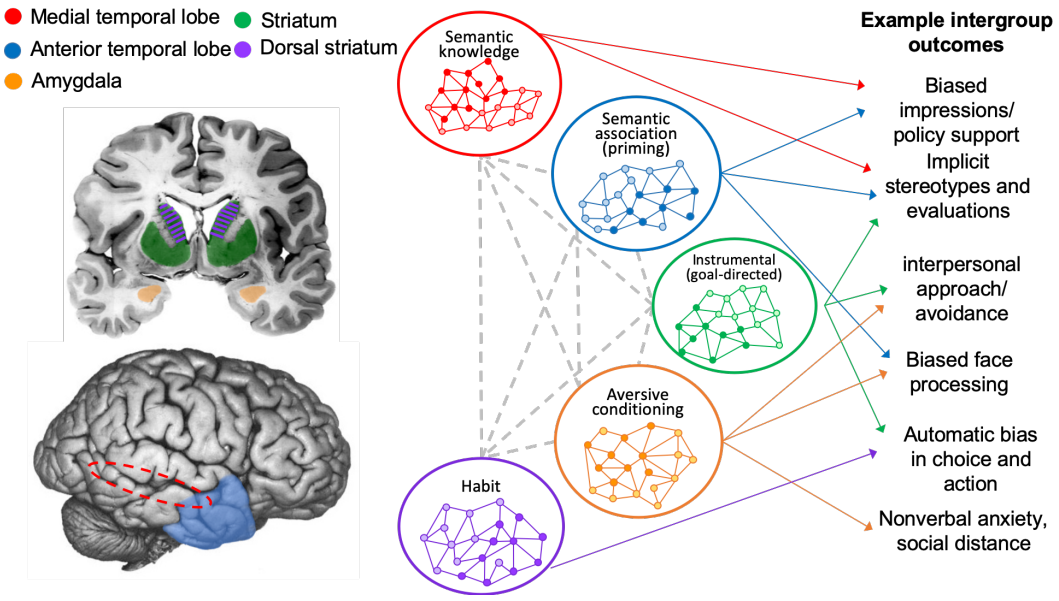

Figure 1. A model of the learning and memory systems through which different forms of intergroup bias are acquired and represented, illustrating their putative neural substrates and examples of their respective intergroup outcomes. Adapted from Amodio (2019).

Stereotypes and conceptual evaluations: The role of semantic memory

Stereotypes represent the conceptual attributes linked to a group, including characterizations of a group's social status and economic standing (e.g., wealthy or poor) and the traits of its members (e.g., smart or hostile). The particular content of a group stereotype may vary between cultures or societies, but what they share in common is a basis in conceptual, or semantic, memory. As such, the process of 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).

In the brain, the process of stereotyping is assumed to involve the same cortical structures that support general forms of semantic memory, including object memory, retrieval, and conceptual activation, such as the temporal lobes and inferior frontal regions (Ralph et al., 2017). Social knowledge, about both people and groups, has been specifically linked to anterior temporal lobe (ATL), including the temporal pole (e.g., Olson et al., 2013; Zahn et al., 2007). Hence, stereotypes and evaluative (i.e., good/bad) conceptual associations—to the extent they represent a social form of semantic processing—should also be associated with activity in these regions.

To date, fMRI studies of stereotyping have largely supported this view. Research by Gilbert et al. (2012) examined neural activity while participants judged Black and White faces according to either a stereotype (athleticism) or an evaluation (potential friendship). To probe stored representations of stereotypes and evaluations, the authors employed multi-voxel pattern analysis (MVPA), which detects spatial patterns of neural activity in fMRI data that differentiate between experimental conditions. Using MVPA, the authors looked for patterns of neural activity during racial judgments of stereotype traits, as opposed to evaluations, that corresponded with participants' scores on separate implicit association tests (IATs) of racial stereotyping and evaluation, respectively. The authors found one region in which activity corresponded to both implicitly-measured stereotyping and implicitly-measured evaluation and correlated, respectively, with the strength of participants' stereotyping and evaluative associations: the ATL. That is, when subjects made trait judgements, stereotyping IAT scores were associated with a pattern of ATL activity that predicted racial differences in stereotype use; when participants

made evaluative judgements, evaluative IAT scores were associated with a different pattern of ATL activity that predicted racial differences in friendship judgments. These findings identified a semantic memory basis for stereotyping, as well as for conceptual evaluative associations, in the ATL.

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 gender stereotype associations on an implicit task (Gallate 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 (Bartholow & Amodio, 2009).

Whereas semantic information about groups, including stereotypes and conceptual evaluations, are stored in the ATL, research suggests that this information is activated and represented in the mPFC when making relevant social judgments (Amodio, 2014). This effect is supported by anatomical connectivity between the ATL and mPFC (de Schotten et al., 2012), and further consistent with the role of the mPFC in representing trait information about individuals during social judgment tasks (Amodio & Frith, 2006; Mitchell et al., 2002) and in stereotypic judgments of gender (Contreras et al., 2012; Quadflieg et al., 2009). In line with this model, Gilbert et al. (2012) found that the application of stereotypes to Black, as opposed to White, target individuals was predicted by patterns of neural activity in the mPFC. By contrast, evaluative judgments of Black, compared with White, target individuals were predicted by neural

patterns in the OFC, a region that also receives strong input from the ATL but is typically involved in evaluative decision-making (O'Doherty et al., 2017).

In summary, the neural basis of stereotyping remains understudied relative to work on prejudiced attitudes and emotion, yet existing research consistently identifies the ATL as supporting the representation of group stereotypes and conceptual evaluative associations. Both kinds of associations may reflect semantic knowledge, and their basis in the ATL is consistent with the broader role of this region in supporting semantic memory. During the process of intergroup decision-making, stereotype knowledge in the ATL is activated and represented in the mPFC, where it guides social judgments.

An affective basis of prejudice? The role of Pavlovian aversive conditioning

Prejudice is often experienced as an affective state, characterized by feelings of fear, threat, or disgust, and this response may occur independently of the semantic associations that characterize stereotypes or conceptual evaluations. Social neuroscience research on prejudiced affect has examined the role of the amygdala as a potential substrate of intergroup threat and fear responses, as well as neural correlates of other emotions such as disgust, linked to the insula (Amodio, 2014). However, the amygdala has received special attention because of its rapid response to a threat and its role in Pavlovian aversive conditioning (LeDoux & Hoffman, 2018), features that may plausibly support an automatic form of prejudice that is affective in nature and that may be learned and expressed independently of semantic processes like stereotypes and conceptual evaluations. 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 (Amodio & Devine, 2005; Amodio et al., 2003). Like most other animals, human acquire fear-conditioned responses to stimuli (Delgado et al., 2006), including human faces (Öhman & Dimberg, 1978), and thus, in theory, this mechanism could also support learned aversions to groups. Although a conditioning basis for prejudice has been suggested by some research showing facilitated association of aversive stimuli with racial outgroup members (e.g., Olsson et al., 2005), to our knowledge, the hypothesis that social prejudice can be formed through Pavlovian aversive conditioning has not yet been tested directly.

Second, aversive conditioning is expressed in a characteristic pattern of behavior that involves freezing, vigilance, and avoidance. Thus, if a form of prejudice were based in aversive conditioning, it would predict a similar pattern of behavior in human intergroup interactions, marked by anxiety and social distancing. Indeed, such 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). An aversive conditioning basis of prejudice may also explain why intergroup anxiety amplifies negative racial evaluations but not stereotyping prior to an interracial interaction (Amodio & Hamilton, 2012). These patterns of intergroup behavior may reflect an aversive conditioning form of prejudice, presumably rooted in amygdala function, that may function differently than judgments and behaviors based in semantic memory processes.

It is notable, however, that despite the existence of Pavlovian aversive conditioning in humans and its potential role in nonverbal and affective expressions of prejudice, neuroimaging evidence for a stronger amygdala response to racial outgroup members has been mixed, at best (Chekroud et al., 2014). Indeed, many fMRI studies of race perception do not report a difference in amygdala response to viewing racial outgroup compared with ingroup members (see Amodio & Cikara, 2021; Mattan et al., 2018). Of those that did, race effects were observed only under specific conditions: for example, following very brief face presentations (Cunningham et al., 2004), when participants made superficial rather than individuating judgments (Wheeler & Fiske, 2005), or in response to direct but not averted gaze (Richeson et al., 2008). However, such studies, like many other early fMRI studies, used very small samples and analytical methods that are no longer considered best practice (Button et al., 2013), and thus such findings should be interpreted with caution.

Other research suggests a more complex role of the amygdala in intergroup responses, such that it primarily guides attention to race, based on its motivational relevance, perhaps especially in situations of threat or anxiety, rather than simply serving to alert threat (Amodio, 2014; Amodio & Cikara, 2021). This perspective reflects findings from studies using alternative measures of amygdala activity, such as the emotion-modulated startle eyeblink response, which historically focused on attentional and motivational aspects of aversive conditioning (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 also 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 (e.g., Brown et al., 2006; Vanman et al., 2013; Van Bavel et al., 2008). This nuanced set of findings suggest that the amygdala response to racial outgroup members may reflect attention driven by social goals and concerns, rather than a simple threat-related affective response to an outgroup member.

In summary, Pavlovian learning may contribute to a specific aspect of prejudice—one that can operate automatically, is associated with negative affect, and is expressed in nonverbal behaviors such as freezing and social distancing. Although reliable fMRI evidence for the amygdala response to race is lacking, the amygdala as a basis for prejudice formation remains plausible, though its specific role in the manifestation of intergroup bias may reflect attentional and motivation processes in addition to affect.

Intergroup bias through social interaction: The role of instrumental learning

Most psychological studies of intergroup bias formation have examined indirect experiences with others, in which we learn about group members by reading descriptions of them or observing their behaviors. Yet much of real-life social behavior involves direct interaction, through actions and feedback between people in a social exchange. Recent social neuroscience findings suggest that this kind of social impression formation 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 intergroup bias, 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, 2011). 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 intergroup bias may help to understand aspects of more automatic forms of bias—particularly those expressed via action, or discriminatory behavior, 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 intergroup bias. Predictions such as these, based on the emerging literature on instrumental learning in social cognition, in combination with new methods borrowed from behavioral economics (Tyler & Amodio, 2015), are currently guiding a new wave of research on the social neuroscience of intergroup bias.

Habits: A basis for automatic intergroup bias?

Automatic forms of intergroup bias are often likened to habits: they appear to emerge from repeated negative experiences with outgroup members, unfold without intention, and resist change (Devine, 1989). Although the concept of "habit" provides an intuitive analogy, what is the evidence that intergroup bias can actually operate like a habit?

Habits typically emerge from instrumental learning—responses that begin as goal-directed reward-driven actions and, over time and repetitions, become routinized as automatic responses that persist irrespective of reward (Robbins & Costa, 2017; Wood & Neal, 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 intergroup bias, 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; Wood, 2017). 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. Nonetheless, a

consideration of this learning mechanism promises to inform our understanding of how implicit bias is expressed and potentially reduced.

Interactive neural system effects in intergroup responses

So far, we have described distinct neurocognitive processes posited to underlie different aspects of intergroup bias, emphasizing their unique neural substrates and psychological characteristics. However, normal human social behavior is complex, and intergroup responses may often involve a combination of multiple memory systems working together. In a social interaction, for example, we may rapidly generate initial trait impressions of a person based on their appearance, learn from their reactions to our own actions, respond affectively to their behavior, and encode the entire experience in semantic and episodic memory—multimodal sources of information recorded in separate but interacting systems. Information from these unique memory systems may then converge, in higher-order representations in the mPFC or OFC, for example, to produce intricate social judgments and behavior. Indeed, neural systems of learning and memory are assumed to operate in concert to support complex cognition and behavior in humans and non-human animals (Packard & Goodman, 2013; Poldrack & Foerde, 2008; Squire, 2004). A consideration of these interactions promises new mechanistic accounts for how different types of intergroup bias may be learned, influence each other, and influence judgments and behaviors.

Recent studies have begun to demonstrate the utility of interactive memory system accounts. For example, research by Hackel and colleagues (Hackel et al., 2015, 2020) demonstrated that humans simultaneously make inferences of both trait concepts and reward value during instrumental social interactions, and that while these complementary streams of

information support different cognitive functions, they are integrated in the OFC to guide decision making. Moreover, a large body of evidence has documented interactions between Pavlovian and instrumental learning systems, both in humans and non-human animals, in which associations learned through Pavlovian conditioning influence otherwise independent instrumental responses (Dickinson & Balleine, 1992; Talmi et al., 2008). A similar phenomenon has been demonstrated in intergroup decision-making, in which racial outgroup-based social threats encoded by a Pavlovian system influence downstream instrumental social responses (Lindström et al., 2015; see also Lindström et al., 2014). By identifying the interactions between systems supporting separable components of intergroup bias, a memory systems approach—rooted in cognitive and social neuroscience—promises to address a broader range of complex intergroup responses than existing sociocognitive accounts.

The interface between individual and systemic bias

An important new direction for intergroup researchers concerns the interplay between systemic bias and individual-level social cognition and behavior. Social psychologists and neuroscientists have focused nearly exclusively on individual-level bias, despite the multi-level nature of discrimination (Trawalter et al., 2020). Yet individual minds are indelibly shaped by the systems in which they operate. Recent research has begun to demonstrate how neural systems can mediate the link between intergroup bias and systemic features, such as scarcity in the economy (Krosch & Amodio, 2019) and socioeconomic disparities (Mattan et al., 2018). To more fully capture the neuroscience of intergroup bias, this field must strive to understand how individual minds contribute to disparities in the context of historical and structural forces.

Conclusion

The memory systems model of intergroup bias represents an integration of social neuroscience research on how individual-level intergroup biases are learned, represented in the mind, and expressed in behavior. This neuroscience framework advances our understanding of traditional intergroup concepts, such as stereotypical beliefs, negative affect, and discriminatory actions, moving beyond prior single- and dual-system sociocognitive accounts, while providing new information about their functions and interactive effects and inspiring new approaches to bias reduction.

References

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience*, *15*, 670-682.

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

Amodio, D. M., & Berg, J. J. (2018). Toward a multiple memory systems model of attitudes and social cognition. *Psychological Inquiry*, *29*, 14-19.

Amodio, D. M., & Cikara, M. (2021). The social neuroscience of prejudice. *Annual Review of Psychology*, *72*, 439-469.

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. (2005). Changing prejudice: The effects of persuasion on implicit and explicit forms of race bias. In T. C. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (2nd ed., pp. 249–280). Thousand Oaks, CA: SAGE Publications.

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

Bartholow, B. D., & Amodio, D. M. (2009). Using event-related brain potentials in social psychological research: A brief review and tutorial. In E. Harmon-Jones & J. S. Beer (Eds.), *Methods in social neuroscience* (pp. 198-232). New York, NY: Guilford Press.

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

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.

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

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365-376.

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.

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.

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 Schotten, M. T., Dell'Acqua, F., Valabregue, R., & Catani, M. (2012). Monkey to human comparative anatomy of the frontal lobe association tracts. *Cortex*, *48*, 82-96.

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

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5-18.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22, 1-18.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*, 62-68.

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.

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

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (357–411). New York, NY: McGraw-Hill.

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.

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

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.

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*, 2592.

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*, *88*, 103948.

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.

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.

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

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.

Lindström, B., Golkar, A., & Olsson, A. (2015). A clash of values: Fear-relevant stimuli can enhance or corrupt adaptive behavior through competition between Pavlovian and instrumental valuation systems. *Emotion*, *15*, 668-676.

Lindström, B., Selbing, I., Molapour, T., & Olsson, A. (2014). Racial bias shapes social reinforcement learning. *Psychological Science*, *25*, 711-719.

Markowitsch, H. J. (1998). Differential contribution of right and left amygdala to affective information processing. *Behavioural Neurology*, *11*, 233-244.

Mattan, B. D., Wei, K. Y., Cloutier, J., & Kubota, J. T. (2018). The social neuroscience of race-based and status-based prejudice. *Current Opinion in Psychology*, 24, 27-34.

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.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences*, *99*, 15238-15243.

Molenberghs, P. (2013). The neuroscience of in-group bias. *Neuroscience & Biobehavioral Reviews*, *37*, 1530-1536.

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

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.

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.

Packard, M. G., & Goodman, J. (2013). Factors that influence the relative use of multiple memory systems. *Hippocampus*, *23*, 1044-1052.

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.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*, 42-55.

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.

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

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

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*, 530-544.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*, 171-177.

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

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human Pavlovian–instrumental transfer. *Journal of Neuroscience*, *28*, 360-368.

Trawalter, S., Bart-Plange, D. J., & Hoffman, K. M. (2020). A socioecological psychology of racism: Making structures and history more visible. *Current Opinion in Psychology*, *32*, 47-51.

Tyler, T. R., & Amodio, D. (2015). Psychology and economics: Areas of convergence and difference. In G. R. Fréchette & A. Schotter (Eds.), Handbook of experimental economic methodology (pp. 181-196). Oxford, UK: Oxford University Press.

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.

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.

Wood, W. (2017). Habit in personality and social psychology. *Personality and Social Psychology Review, 21*, 389-403.

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

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*, 6430-6435.