Social Cognitive and Affective Neuroscience, 2016, 1–12

doi: 10.1093/scan/nsw147
Original article

Motivation alters impression formation and related neural systems

Brent L. Hughes,[1,2] Jamil Zaki,[1] and Nalini Ambady[1,†]

[1]Department of Psychology, University of California, Riverside, CA 92521, USA and [2]Department of Psychology, University of California Riverside, Los Angeles, CA, USA

Correspondence should be addressed to Brent L. Hughes, Department of Psychology, University of California, 900 University Ave., Riverside, CA 92521, USA. E-mail: blhughes@stanford.edu

[†]In Memoriam

Abstract

Observers frequently form impressions of other people based on complex or conflicting information. Rather than being objective, these impressions are often biased by observers' motives. For instance, observers often downplay negative information they learn about ingroup members. Here, we characterize the neural systems associated with biased impression formation. Participants learned positive and negative information about ingroup and outgroup social targets. Following this information, participants worsened their impressions of outgroup, but not ingroup, targets. This tendency was associated with a failure to engage neural structures including lateral prefrontal cortex, dorsal anterior cingulate cortex, temporoparietal junction, Insula and Precuneus when processing negative information about ingroup (but not outgroup) targets. To the extent that participants engaged these regions while learning negative information about ingroup members, they exhibited less ingroup bias in their impressions. These data are consistent with a model of 'effortless bias', under which perceivers fail to process goal-inconsistent information in order to maintain desired conclusions.

Key words: motivated cognition; impression formation; intergroup processes; social cognition; cognitive control

People contain multitudes, in that they often exhibit complex and even conflicting behaviors. For instance, the same person might behave morally in one case and immorally in another. Despite this, observers are able to form quick and stable impressions of the people around them (Ambady *et al.*, 2000). How do observers accomplish this feat? Classic and contemporary research suggests that observers integrate over the social information they encounter, and update their impressions according to several stable 'rules' (Asch, 1946; Anderson, 1965; Aronson *et al.*, 1966; Reeder and Brewer, 1979; Fiske, 1980; Skowronski and Carlston, 1989; Freeman and Ambady, 2011; Zaki, 2013). For instance, observers weigh information learned first more heavily than subsequently learned information (e.g. Asch, 1946). Observers also weigh negative information more heavily than positive information, such that they form negative impressions about social targets after hearing about equal proportions of positive and negative behaviors those targets produce (e.g. Hamilton and Zanna, 1972; Reeder and Brewer, 1979; Fiske, 1980; Skowronski and Carlston, 1989; Baumeister *et al.*, 2001; Knobe, 2003). For instance, after learning about someone who donates money to charity but also sells fake memorabilia on Ebay, observers will tend to form negative (rather than neutral) impressions of such a target.

Such updates are associated with activity in a broad network of brain regions including lateral prefrontal cortex (LPFC), and medial prefrontal cortex (MPFC) including the dorsal anterior cingulate cortex (dACC), temporoparietal junction (TPJ), posterior cingulate (PCC) and precuneus, among other regions (Schiller *et al.*, 2009; Freeman *et al.*, 2010; Zaki *et al.*, 2010; Baron *et al.*, 2011; Cloutier *et al.*, 2011; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013a,b; Hackel *et al.*, 2015; Stanley, 2015). Activation in these regions while observers encode new information tracks the magnitude of impression updating, or the extent to which observers modify their initial impressions

Received: 12 June 2015; **Revised:** 23 August 2016; **Accepted:** 4 October 2016

based on subsequent information. Importantly, neuroscientific work in this domain has examined brain activity associated with updating impressions based on a combination of positive and negative information (e.g. Baron *et al.*, 2011; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013b); as we describe above, such cases normatively result in worsened impressions of targets. As such, activation in LPFC, MPFC and dACC tracks the formation of more negative impressions that observers typical form after learning mixed information about targets (Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013b).

However, observers should be reluctant to worsen their impressions in some cases. In particular, observers often want to see certain individuals (e.g. friends) in a positive light, even when faced with mixed evidence (e.g. Brewer, 1999; Taylor and Brown, 1988). In order to maintain favorable opinions about their friends or fellow group members, observers frequently favor goal-consistent data that supports their views over goal-inconsistent data that refutes their views (Tajfel and Turner, 1979; Brewer, 1999; Dovidio and Gaertner, 2010). For instance, people expect that ingroup members will exhibit more positive and fewer negative behaviors than outgroup members, remember more positive and less negative information about ingroup, as compared with outgroup members (Howard and Rothbart, 1980; Foddy *et al.*, 2009), and use more positive language in describing the same positive behaviors exhibited by ingroup, as compared with outgroup members (Maass *et al.*, 1989). These cases exemplify the broader phenomenon of motivated cognition, which describes the myriad ways through which goals and needs shift people's thinking, allowing them to arrive at their preferred conclusions (Kunda, 1990; Dunning, 2014).

Recent neuroscientific work identifies a consistent pattern of neural activation associated with motivated cognition (Hughes and Zaki, 2015). Crucially, motivated effects on cognition often track 'reduced' activity in regions associated with evaluation, including LPFC, dACC and MPFC (Egner, 2009; Braver, 2012; Roy *et al.*, 2012; Shenhav *et al.*, 2014). These reductions in activity further track the extent to which people ignore goal-inconsistent information that threatens positive self-views (Krusemark *et al.*, 2008; Somerville *et al.*, 2010; Sharot *et al.*, 2011), or their view of favored political candidates (Kato *et al.*, 2009), and also the extent to which they globally enhance their perceptions of themselves or close others (Beer and Hughes, 2010; Hughes and Beer, 2012a,b). Engagement of these neural systems is associated with reduced bias, for instance, reducing the extent to which people ignore goal-inconsistent information and see themselves and close others as superior to their peers. The consistent inverse relationship between neural systems of evaluation and social cognitive biases suggests a model of 'effortless bias' in motivated cognition (Hughes and Zaki, 2015). Motivated biases exhibit other characteristics of automaticity, for instance arising quickly (Balcetis and Dunning, 2006; Mulder *et al.*, 2012), without a perceiver's awareness (Pronin 2007; West *et al.*, 2012), and even when cognitive resources are limited (Lench and Ditto, 2008; Beer and Hughes, 2010; Beer *et al.*, 2013). Perceivers may quickly and effortlessly anchor a judgment on a biased starting point or prior belief, unless they exert effort or are explicitly motivated to correct for their biases (Hughes and Beer, 2012b). In the context of impression updating, perceivers may effortlessly favor goal-consistent information while failing to appropriately account for goal-inconsistent information.

Although nascent neuroscientific work examines motivated cognition and impression updating separately, very little research examines the intersection of these key social cognitive processes. Here, we examine whether and how motivated cognition biases impression formation and its related neural systems in an intergroup context. In particular, observers should have no problem worsening their impression of outgroup members in response to unsavory information, because such information is not inconsistent with the goals of observers. In contrast, they might fail to worsen impressions about ingroup members based on such information, because doing so runs counters to observers' goals to see ingroup members in a positive light. Further, this bias should track observers' failures to engage neural systems associated with impression change when they encode goal-inconsistent information, such as negative behaviors of ingroup members (e.g. LPFC, MPFC, dACC, TPJ, PCC/Precuneus: Schiller *et al.*, 2009; Cloutier *et al.*, 201b; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013a,b; Hackel *et al.*, 2015; Stanley, 2015). Finally, observers' engagement of those same neural systems when encoding negative information should be associated with reduced ingroup bias.

Here, we used functional magnetic resonance imaging (fMRI) to examine these predictions. Stanford undergraduates formed initial impressions of social targets, and then learned that targets were either fellow Stanford students (ingroup), University of California at Berkeley students (outgroup), or neither (control). Participants were then exposed to positive and negative information about each target and asked to update their impression based on that information. Consistent with a model of effortless bias emerging from the neuroscience of motivated cognition (Hughes and Zaki, 2015), we expected that people would exhibit an effortless intergroup bias when encoding new information. In particular, although they might integrate positive and negative information about outgroup targets equally we predicted that they would fail to update impressions of ingroup members based on negative information. We further predicted that this failure would be associated with reduced activity in a system of brain regions associated with impression updating while observers encode new information, consistent with a model of effortless bias.

Methods

Participants

Twenty-six participants (15 female, mean age = 19.1 years, SD = 1.1) were recruited in compliance with the human subjects regulations of Stanford University and compensated with $15/h or course credit (15 female, mean age = 19.1 years, SD = 1.1). Five participants were excluded from analyses (three due to excessive head motion, one due to scanner malfunction, one for responding to <50% of trials). The remaining 21 participants (12 female, mean age = 18.8 years, SD = 0.75) were all right-handed, native English speakers, free from medications and psychological and neurological conditions, and had normal or corrected-to-normal vision. All participants were prescreened to ensure that they were Stanford University students, the ingroup in our study. Finally, participants completed the Collective Self-Esteem Scale (Luhtanen and Crocker, 1992) to ensure that they would experience a motivation to favor ingroup members. Participants all reported positive associations with their social identity as Stanford students (M = 5.5, SD = 0.4).

Procedure

Participants completed a social learning task in which they formed impressions of ingroup members (i.e. other Stanford University students) or outgroup members (i.e. UC-Berkeley

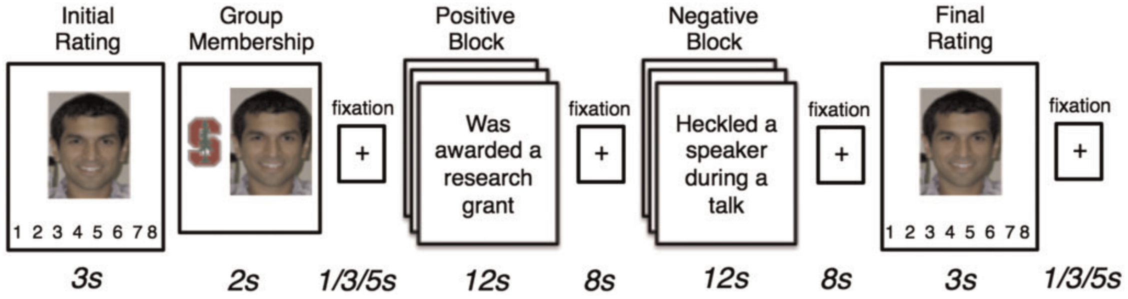

Fig. 1. Example ingroup trial. On each trial, participants had 3 s to form an initial impression of a social target. Participants were then exposed to the group membership of the social target (2 s). Participants then read a block of positive information (three positive behaviors for 12 s total) and a block of negative information (three negative behaviors for 12 s total), presented in counterbalanced order. Finally, participants had 3 s to provide an updated impression of the social target.

students). In addition to the ingroup and outgroup conditions of interest, participants also formed impressions of social targets devoid of any group membership information (control trials). Each trial began with a First Impression slide consisting of a forward facing facial photograph and a rating scale (1–8), during which participants were asked to form an initial impression of likability (3 s). Participants were then presented with the Group Membership of the target (Stanford logo, Cal logo; 2 s), or a Fixation cross for Control targets (2 s). Participants then read one block of positive information (three positive behaviors presented serially for 4 s each) and one block of negative information (three negative behaviors presented serially for 4 s each). Encoding blocks of positive and negative information were presented in counterbalanced order within the group membership condition. Finally, participants were presented with the forward facing facial photograph of the target and rating scale (1–8), and asked to provide a Final Impression of likability based on the information they had learned (3 s). Phases within trials were separated by a randomly jittered inter-trial interval (see Figure 1). Trial order was randomized within runs, with an equal number of trials per group condition per run.

Participants completed 4 functional runs of 12 trials per run for a total of 48 trials (16 each for ingroup, outgroup and control). Visual stimuli were presented using E-prime and projected onto a large-screen flat-panel display monitor that participants viewed in a mirror mounted on the scanner.

Stimuli

Social targets were represented by photographs of faces, and group membership information was depicted by a Stanford University logo (ingroup condition), Cal logo (outgroup condition), or control (no group membership displayed). Photographs were drawn from the first author's photo database and consisted of color pictures of forward-looking male faces with neutral expressions. Photographs were randomly distributed to belong to the Ingroup, Outgroup and Control conditions and were equated across White, Asian, Hispanic and Black faces. As this experiment was not designed to test whether demographic variables interact with manipulated group membership, we lack statistical power to systematically examine these interactions. Future work should examine the interaction between preexisting demographic differences and assigned group membership in impression formation processes.

Positive and negative information about each social target consisted of sentences describing behaviors previously rated on valence (on a 0–10 scale: Positive sentences: $M = 7.26$, $SD = 1.33$; Negative behaviors: $M = 3.1$, $SD = 1.12$; Fuhrman *et al.*, 1989). Each social target was paired with three positive and three negative behaviors, presented in counterbalanced order within each group membership condition. Positive and negative behaviors were randomly assigned to ingroup, outgroup and control conditions and equated on valence to ensure that social targets and group membership conditions did not systematically differ on information valence. The blocks of three positive and three negative behaviors were pilot tested to ensure that participants had ample time to read each sentence.

MRI data acquisition

All images were collected on a 3.0 T GE Discovery MR750 scanner at the Center for Cognitive and Neurobiological Imaging at Stanford University. Functional images were acquired with a T2*-weighted gradient echo pulse sequence (TR = 2 s, TE = 24 ms, flip angle = 77) with each volume consisting of 46 axial slices (2.9-mm-thick slices, in-plane resolution 2.9-mm isotropic, no gap, interleaved acquisition). Functional images were collected in four runs (consisting of 12 trials per run, 48 trials total). High-resolution structural scans were acquired with a T1-weighted pulse sequence (TR = 7.2 ms, TW = 2.8 ms, flip angle = 12) after functional scans, to facilitate their localization and coregistration.

MRI data analysis

All statistical analyses were conducted using SPM8 (Wellcome Department of Cognitive Neurology). Functional images were reconstructed from *k*-space using a linear time interpolation algorithm to double the effective sampling rate. Image volumes were corrected for slice-timing skew using temporal sinc interpolation and for movement using rigid-body transformation parameters. Functional data and structural data were coregistered and normalized into a standard anatomical space (2-mm isotropic voxels) based on the echo planar imaging and T1 templates (Montreal Neurological Institute), respectively. Images were smoothed with a 5-mm full-width at half-maximum Gaussian kernel. To remove drifts within sessions, a high-pass filter with a cutoff period of 128 s was applied. Visual inspection of motion correction estimates confirmed that no subject's head motion exceeded 3.0 mm in any dimension.

Our primary analytic strategy focused on brain activity while participants encoded new information about each social target. In particular, we sought to isolate neural structures in which activation during encoding predicted subsequent changes in impression about targets. Our primary question of interest directly relates to how observers dynamically encode (or fail to encode) new information if it clashes with a desire to maintain favorable impressions of ingroup members. To address this question, we were interested in neural activation during encoding that predict subsequent changes in impression, and how this neural activation at encoding may be modulated by group membership. To that end, we examined neural activation during encoding that was parametrically modulated by the degree of subsequent impression change, that is, the difference between final and initial impression rating.

The GLM consisted of 20 regressorso of interest and 6 regressors of noninterest. Three regressors modeled the First Impression—based on photographs—of ingroup targets, outgroup targets and control targets before the presentation of group membership (modeled as stick functions with 0-s duration). Two regressors modeled the Group Affiliation logos (modeled as stick functions with 0-s duration). Three regressors modeled the Final Impression of these three target types after participants learned information about them (modeled as stick functions with 0-s duration). Finally, three regressors of interest modeled, more specifically, blocks in which participants learned Positive information (three regressors) and Negative information (three regressors) about each target type. These regressors were modeled as boxcar functions with a 12-s duration. We included parametric modulators reflecting the degree to which participants updated their impressions of each target type for the Positive (three parametric regressors) and Negative Encoding blocks (three parametric regressors). This allowed us to test for neural responses during the encoding of Positive and Negative information that parametrically tracked subsequent changes in impression (Final—First impression ratings). A positive sign on this impression change regressor indicates that impressions became more positive from initial to final impression. A negative sign on this impression change regressor indicates that impressions worsened, or became more negative from initial to final impression. Six regressors of noninterest modeled participant head movement during scans. The regressors of interest were convolved with a canonical (double-gamma) hemodynamic response function.

We employed this model for three main analyses. First, we computed a contrast weighted positively on parametric modulators for all of the Encoding blocks, regardless of group membership, 'Encoding blocks' > 'Fixation baseline' (where 'Fixation baseline' represents the unmodeled Fixation epochs to serve as a baseline condition). This allowed us to isolate neural activity during encoding that predicted subsequent changes in impressions, regardless of targets' group membership or valence of encoded information (positive or negative information). Positive relationships between brain activation and parametric changes in impression would indicate that, as brain activation increases, impressions become more favorable from initial to final impression as a function of the encoded information. Negative relationships between brain activation and parametric changes in impression indicate that, as brain activation increases, impressions become more negative from initial to final impression as a function of the encoded information.

Importantly, we predicted that observers would be especially motivated to ignore information—particularly negative information—about ingroup members. We predicted that reduced brain activity during encoding about ingroup members would predict failure update ingroup impressions, consistent with a model of 'effortless bias' in motivated cognition. To this end, we contrasted parametric modulators for 'Ingroup Encoding blocks *vs* Outgroup Encoding blocks and Ingroup Negative Encoding blocks *vs* Outgroup Negative Encoding blocks'. These contrasts examine whether group membership significantly affects the relationship between brain activation and impression change. Therefore, negative differences between Ingroup and Outgroup Encoding conditions would indicate that brain activation more strongly tracks worsened impressions for Ingroup than for Outgroup targets.

Finally, we examined parametric contrasts for the 'Ingroup Encoding blocks' > 'Fixation baseline' and 'Outgroup Encoding blocks' > 'Fixation baseline'. These contrasts address not whether brain regions 'differentially' track impression changes during Encoding between Ingroup and Outgroup conditions, but rather whether brain activation significantly tracks impression changes within each group condition *vs* baseline. Positive relationships would indicate that, as brain activation increases, impressions become more favorable from initial to final impression as a function of the encoded information. Negative relationships indicate that, as brain activation increases, impressions become more negative from initial to final impression as a function of the encoded information.

Main effect maps were thresholded at $P < 0.005$, with a spatial extent threshold of $k = 23$, corresponding to a threshold of $P < 0.05$ corrected for multiple comparison (derived from the latest release of the AFNI program 3dClustSim).

ROI analysis

We conducted an additional GLM exclusively to visualize the effects from the continuous, trial-by-trial parametric analyses reported earlier. First, trials were binned into three types: (i) trials for which impressions became more positive, (ii) trials for which impressions did not change and (iii) trials for which impressions became more negative. The number of trials in each bin varied by participant, as the trial-by-trial analysis was idiosyncratically defined.

The GLM consisted of 14 regressors of interest and 6 regressors of noninterest. Three regressors modeled the First Impression—based on photographs—of ingroup targets, outgroup targets and control targets before the presentation of group membership (modeled as stick functions with 0-s duration). Two regressors modeled the Group Affiliation logos (modeled as stick functions with 0-s duration). Three regressors modeled the Final Impression of these three target types after participants learned information about them (modeled as stick functions with 0-s duration). Finally, three regressors of interest modeled, more specifically, blocks in which participants encoded Positive information (three regressors) and Negative information (three regressors) about each target type. These regressors were modeled as boxcar functions with a 12-s duration. Six regressors of noninterest modeled participant head movement during scans. The regressors of interest were convolved with a canonical (double-gamma) hemodynamic response function.

Parameter estimates were extracted from the 3 main contrasts reported below: (i) Encoding blocks > Fixation, regardless of group membership, (ii) Ingroup Encoding blocks > Outgroup Encoding blocks and (iii) Ingroup Negative Encoding blocks > Outgroup Negative Encoding blocks. Parameter estimates were extracted from significant activation clusters in each respective contrast from regions previously found to be associated with impression updating (Schiller *et al.*, 2009; Cloutier *et al.*, 2011b; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013a,b; Hackel *et al.*, 2015; Stanley, 2015). We plot parameter estimates from LPFC and dACC for visualization purposes alone, as these regions are (i) consistently activated across all of our reported analyses and (ii)

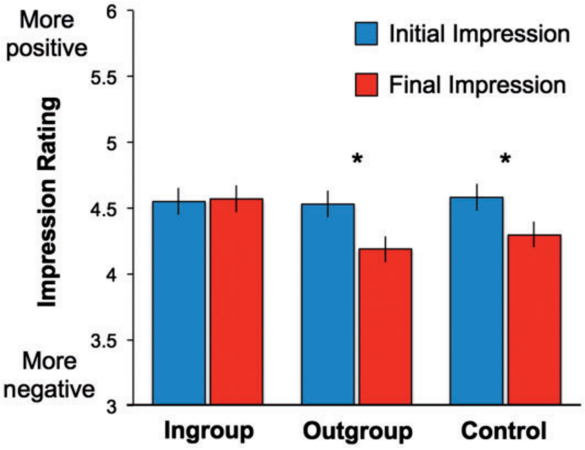

Fig. 2. Impressions varied significantly by group. Whereas impression ratings of outgroup members and control targets became more negative from initial to final impression, impression ratings of ingroup members did not. Error bars represent SEM.

consistently observed to play a role in updating impressions in the face of new information.

Results

Behavioral results

Participants updated their impressions in a way consistent with motivated bias. We found a significant interaction between group membership (ingroup, outgroup and control) and time of impression (initial impression, final impression) in predicting biased impression updates [$F(2,19) = 4.37$, $P < 0.05$; see Figure 2]. After learning positive and negative information about targets, observers rated outgroup and control targets significantly more negatively than they had initially [Outgroup: $t(20) = 2.90$, $P < 0.05$, Control: $t(20) = 2.70$, $P < 0.05$]. However, they showed no such change in their ratings of ingroup members [$t(20) = 0.32$, $P = 0.75$]. As expected, First Impression ratings did not differ as a function of group (as group-related information was presented 'after' the first impressions were made; all ts < 1.20). However, final impressions were significantly more favorable for Ingroup targets than for Outgroup [$t(20) = 2.93$, $P < 0.05$] or Control targets [$t(20) = 2.17$, $P < 0.05$). Final impressions did not differ between Outgroup and Control targets [$t(20) = 1.68$, $P = 0.12$]. Finally, there were no differences in reaction time for Initial or Final rating between any group membership condition, or between Positive and Negative impression ratings (all ts < 1.1).

These findings demonstrate that motivational influences in impression formation are at times characterized by ingroup favoritism rather than outgroup derogation, which underscores the long-standing dissociation between ingroup love and outgroup hate (Brewer, 1999).

Follow-up behavioral results

The behavioral results described earlier leave open the possibility that participants may have worsened their outgroup impressions because of their 'outgroup' label, and not due to differential weighting of negative *vs* positive information. Although participants did differentially weigh negative *vs* positive information when forming impressions of Control targets, the possibility remains that participants may have treated Control targets as Outgroup members. In order to explicitly determine whether participants indeed overweigh negative *vs* positive information when forming impressions, we recruited an independent sample of participants on Amazon Mechanical Turk to complete a follow-up experiment. Participants ($n = 100$) formed an initial impression of likability based on a photograph on a eight-point scale, then read six pieces of behavioral information about the target, and finally updated their impression of likability on the same eight-point scale. Participants were randomly assigned to one of two conditions of the task. In the 'neutral' condition, participants read six pieces of neutral, unvalenced information about the social target. In the 'valenced' condition, participants read three pieces of positive and three pieces of negative information (in counterbalanced order) about the social target. We selected a subset of positive ($M = 6.98$, SD = 0.45) and negative ($M = 3.29$, SD = 0.54) behavioral information used in the fMRI experiment and included additional neutral, unvalenced behavioral information ($M = 5.05$, SD = 0.15) used in previous research (Fuhrman *et al.*, 1989).

We found a significant interaction between Condition (Valenced, Neutral) and Time of Impression, $F(2,97) = 5.82$, $P < 0.05$. Whereas participants significantly worsened their impressions of social targets in the Valenced condition [$t(49) = -3.9$, $P < 0.05$), they did not significantly change their impression in the neutral condition [$t(49) = 0.2$, $P = 0.8$]. These findings add supportive evidence that participants weigh negative information more strongly than positive information as demonstrated in previous work (e.g. Fiske, 1980; Baumeister *et al.*, 2001; Knobe, 2003). These data suggest that, in the presence of equal proportions of positive and negative information, people normatively worsen their impressions. Therefore, a failure to worsen impressions in these contexts should reflect a bias that may be extended to ingroup members and can be used as a proxy of favoritism.

Table 1. Brain activation during encoding that parametrically tracks subsequent impression changes overall.

Region of Activation	BA	x	y	z	T-stat	Cluster size
Encoding Blocks > Baseline						
DACC	24	0	18	38	5.57	167
TPJ/IPL	39/40	48	−46	22	4.76	172
		−56	−48	30	4.42	87
Insula		−56	10	6	4.69	27
		36	14	0	4.21	84
LPFC	44/45	26	44	32	3.46	34
Precuneus	7	−10	−64	28	3.61	26

BA = Brodmann's Area; DACC = dorsal anterior cingulate cortex; TPJ = temporoparietal junction; IPL = inferior parietal lobule; LPFC = lateral prefrontal cotex.

Neuroimaging results

Brain activity during learning predicts subsequent changes in impression overall. Although participants learned information about social targets, activity in a broad network of regions—including dACC, right dlPFC, bilateral IPL/TPJ, bilateral insula and precuneus (see Table 1 and Figure 3)—predicted later changes in participants' impressions about those targets (All Encoding Blocks > Baseline). Importantly, greater activation in these regions parametrically tracked subsequent 'worsening' of impressions from Initial to Final impression (see Figure 4). This is broadly consistent with prior work implicating these regions in updating impressions, especially based on a mix of positive and

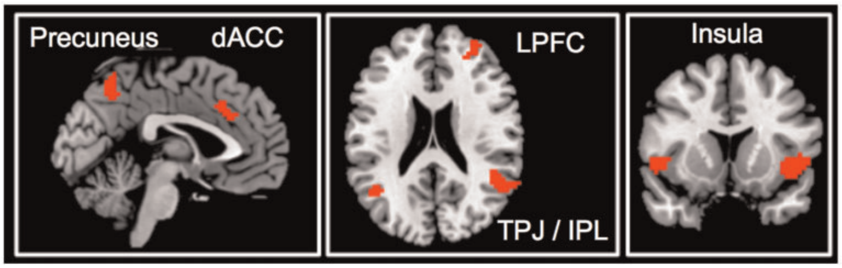

Fig. 3. Parametric analyses revealed that brain activation in dACC and Precuneus ($x = 0$), LPFC and TPJ/IPL ($z = 24$) and Insula ($y = 12$) during encoding predicted subsequent changes in impression overall.

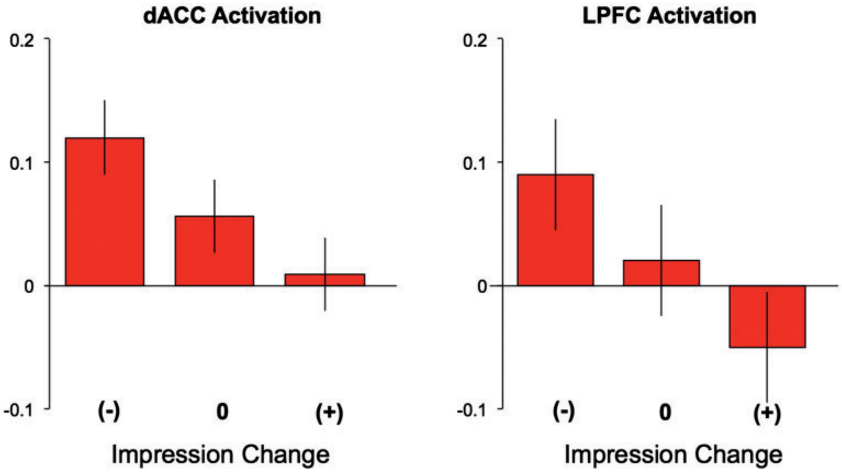

Fig. 4. Activity in dACC and LPFC parametrically tracked worsened impression overall (All Encoding Blocks > Fixation Baseline). Error bars represent SEM.

negative information (e.g. Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013b).

Brain activity that predicts impression change is bounded by group membership. We reasoned that motives would prevent participants from encoding negative information about ingroup members, and that encoding such negative information would require observers to engage neural systems associated with updating impressions. Consistent with this prediction, greater activity in dACC, bilateral dlPFC and bilateral IPL/TPJ while participants learned about ingroup (*vs* outgroup) targets parametrically tracked subsequent worsening of impressions about 'ingroup' targets (Ingroup Encoding Blocks > Outgroup Encoding Blocks; see Table 2 and Figure 5). Follow-up analyses revealed that greater activation in these regions, and additional activation in Insula and Precuneus, significantly tracked subsequent worsening of impressions about ingroup targets (Ingroup *vs* Baseline), but not outgroup targets (Outgroup *vs* Baseline), even at a more liberal $P < 0.05$ uncorrected threshold (see Table 2).

Group membership influences the encoding of negative information. If people fail to specifically encode negative information about ingroup members, then counteracting this bias should engage neural systems associated with impression updating processes. Consistent with this hypothesis, greater activation in bilateral dlPFC, bilateral IPL/TPJ and bilateral insula while participants learned negative information about ingroup (*vs* outgroup) targets predicted subsequent worsening of impressions about ingroup targets (Ingroup Negative Encoding Blocks > Outgroup Negative Encoding Blocks; see Table 3). Follow-up analyses revealed that greater activation in these regions while learning negative information predicted subsequent worsening of impressions about ingroup targets (Ingroup Negative Encoding Blocks *vs* Baseline), but not for outgroup targets (Outgroup Negative Encoding Blocks *vs* Baseline), even at a more liberal $P < 0.05$ uncorrected threshold (see Table 3 and Figure 6). Moreover, marginally significant activation in dACC while learning negative information predicted subsequent worsening of impressions about ingroup targets (*vs* Fixation baseline), but not outgroup targets (*vs* Fixation baseline; see Table 3). Since impression updating was, in general, characterized by ingroup favoritism, this suggests that this system of brain regions plays a corrective role: reducing ingroup biases in impression formation.

Finally, just as people may fail to encode negative information about ingroup members to maintain ingroup favoritism, they may likewise fail to encode positive information about outgroup members. Counteracting this outgroup bias might also engage neural systems associated with impression updating processes. However, we did not observe any significant neural relationships that tracked 'improved' impressions about outgroup targets while learning positive information about them (Outgroup Positive Encoding Blocks *vs* Baseline). Likewise, we did not observe significant neural relationships between improved impressions about ingroup or control targets while learning positive information about them (Ingroup Positive Encoding Blocks *vs* Baseline; Control Positive Encoding Blocks *vs* Baseline).

Discussion

Our data suggest that motivation biases the encoding of information and influences neural systems of impression updating.

Table 2. Brain activation during encoding that parametrically tracks subsequent impression changes for ingroup (but not outgroup) members.

Region of Activation	BA	x	y	z	T-stat	Cluster size
Ingroup Encoding > Outgroup Encoding Contrast						
DACC	24	6	20	22	4.32	98
LPFC	44/45	52	28	32	4.01	35
		−42	30	30	3.66	34
		46	44	8	3.47	45
		−48	8	22	3.59	30
TPJ/IPL	39/40	48	−56	22	3.71	49
		−40	−58	22	3.79	42
Ingroup Encoding > Baseline						
DACC	24	2	24	34	4.09	103
LPFC	44/45	32	44	34	4.54	86
		46	40	20	3.96	28
TPJ/IPL	39/40	50	−48	22	3.52	148
		−44	−58	20	3.95	31
Insula		−40	14	−6	3.98	31
Precuneus	7	0	−60	46	3.74	62
Outgroup Encoding > Baseline						
No significant activation clusters						
Control Encoding > Baseline						
Precentral gyrus	6	−33	5	40	3.26	76
Postcentral gyrus	3	36	−20	43	3.53	208
Medial temporal lobe	48	−48	−16	−2	3.57	33

BA = Brodmann's Area; DACC = dorsal anterior cingulate cortex; TPJ = temporoparietal junction; IPL = inferior parietal lobule; LPFC = lateral prefrontal cotex.

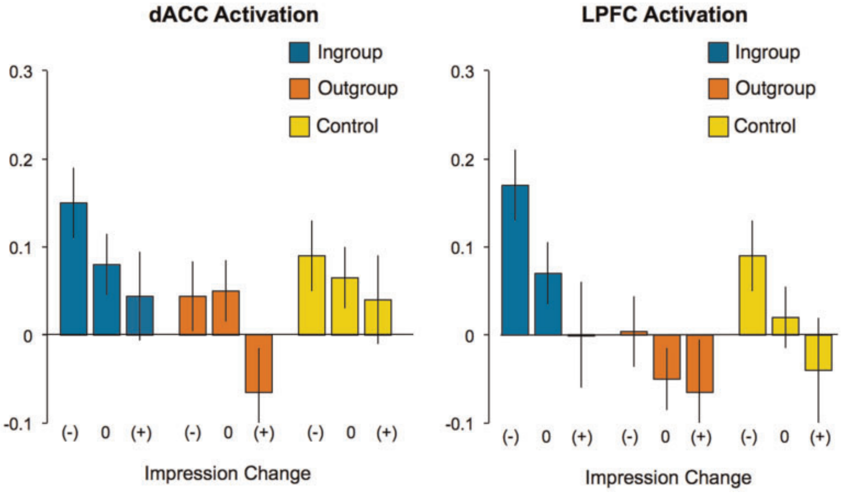

Fig. 5. Activity in dACC and LPFC parametrically tracked worsened impressions of ingroup (but not outgroup) members, mitigating ingroup biases in impression formation (Ingroup Encoding Blocks > Outgroup Encoding Blocks). Error bars represent SEM.

Observers asymmetrically updated impressions about others based on targets' group membership after learning a mix of positive and negative information about them. Whereas impressions of outgroup and control targets worsened, impressions of ingroup targets did not. These findings highlight one mechanism by which observers maintain favorable group impressions. Whereas observers evenly encode negative and positive information about outgroup and control targets, they here failed to encode negative information about ingroup targets, and as such maintained biased, positive impressions of those targets.

Consistent with previous work, we observed a large network of brain regions involved in updating impressions (e.g. Schiller *et al.*, 2009; Cloutier *et al.*, 2011b; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013a,b; Hackel *et al.*, 2015; Stanley, 2015), and especially with worsened impressions of others based on a mix of positive and negative information. Consistent with the model of 'effortless bias' emerging from the neuroscience of motivated cognition (Hughes and Zaki, 2015), failure to engage these regions tracked participants' behavioral ingroup bias, and activation in this system was associated with a reduction of such bias. In particular, activation in LPFC, dACC, IPL/TJP, insula and precuneus preferentially tracked worsened impressions about ingroup (but not outgroup) members. This effect was specifically driven by activity when participants encoded negative information about ingroup members, an information-type observers were likely motivated to avoid. Together, these findings suggest that people heed motives by failing to encode negative information into their impressions, and counteracting this bias engages greater activation in neural systems associated with impression updating.

Table 3. Brain activation while encoding negative information that parametrically tracks subsequent impression changes for ingroup (but not outgroup) members.

Region of Activation	BA	x	y	z	T-stat	Cluster size
Ingroup Negative Encoding > Outgroup Negative Encoding Contrast						
LPFC	44/45	50	28	28	4.50	130
		−44	34	28	3.95	117
TPJ/IPL	39/40	50	−60	24	3.61	76
		−36	−58	22	3.91	66
Insula		−32	18	−11	3.50	23
Parahippocampal cortex		32	−22	−14	3.50	122
Temporal cortex		−45	−12	−14	3.46	33
Ingroup Negative Encoding > Baseline						
LPFC	44/45	52	20	30	4.14	150
		42	50	−4	4.23	25
		−40	24	38	3.41	26
TPJ/IPL	39/40	60	−54	14	4.08	122
		−60	−52	36	4.00	49
		−45	−58	22	3.70	62
Insula		−45	14	−4	4.08	122
		45	20	−8	3.47	82
DMPFC	8/9	8	22	54	3.57	45
DACC	24	2	24	32	3.21	23
Outgroup Negative Encoding > Baseline						
No significant activation clusters						
Control Negative Encoding > Baseline						
No significant activation clusters						

BA = Brodmann's Area; LPFC = lateral prefrontal cortex; TPJ = temporoparietal junction; IPL = inferior parietal lobule; DMPFC = dorsomedial prefrontal cortex; DACC = dorsal anterior cingulate cortex.

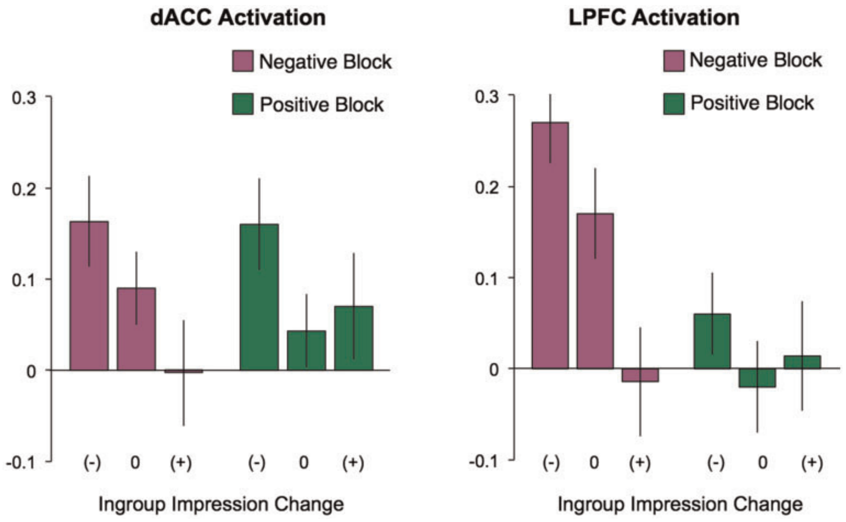

Fig. 6. Activity in dACC and LPFC while learning negative information—an information-type observers are likely motivated to avoid—predicted subsequent worsened impressions about ingroup members (Ingroup Negative Encoding Blocks > Fixation Baseline). Error bars represent SEM.

These findings extend prior work in a number of ways. First, these findings demonstrate that motivation influences brain systems underlying impression updating. Recent work demonstrates that updating impressions when faced with positive and negative information is associated with activation in a broad network of regions, including LPFC, dACC, MPFC, TPJ and PCC/Precuneus (Schiller *et al.*, 2009; Baron *et al.*, 2011; Cloutier *et al.*, 2011b; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013a; Hackel *et al.*, 2015; Stanley, 2015). The present findings conditionalize these insights by demonstrating that the relationship between increased neural activation and impression updates are bounded by group membership. Specifically, activation in LPFC, dACC, TPJ, insula, precuneus and other regions tracked worsened impressions about ingroup, but not outgroup members. Whereas observers might expect—and even enjoy—unflattering information about outgroup members, they likely neither expect nor want to hear such information about ingroup members (Howard and Rothbart, 1980; Brewer, 1999; Foddy *et al.*, 2009). Our data dovetail with prior work in demonstrating that observers appear to ignore or undervalue such undesirable news about ingroup members. We further demonstrate that this motivated avoidance tracks reduction in the activity of brain regions generally associated with impression formation, consistent with an emerging model of 'effortless bias' (Hughes and Zaki, 2015). Finally, activity in this same system tracked a reduction of bias, likely allowing observers to appropriately consider even information that runs counter to their motives, such as unexpected and undesirable information about ingroup members. Following from previous work, we capitalized on trial-by-trial variability to track the neural associations with biased impression updates. As our sample all had very positive associations with their group identity, we were limited in our ability to resolve individual variability in bias. Future work should examine whether individual variability in biased impression updating significantly modulates neural activation in these regions.

Importantly, we observed impression formation and its related biases only in cases where observers learned equal amounts of positive and negative information about targets. This context matches other behavioral and neuroscientific explorations of impression updating (e.g. Schiller *et al.*, 2009; Ma *et al.*, 2012; Bhanji and Beer, 2013; Mende-Siedlecki *et al.*, 2013a,b), but of course does not characterize many instances of impression formation outside the lab. As described earlier, learning 50% positive and 50% negative information about a social target typically results in a negative overall impression of that person. Most people strive to instead emit largely positive cues to others. In such contexts, we might expect the relationship between motivation, impression updating and brain activity to differ. For instance, if given largely positive information about targets, observers might gladly update impressions about ingroup members but fail to appropriately update impressions of outgroup members. In this case, activity in regions associated with impression updating might track appropriate 'positive' updating of outgroup impressions. More generally, we believe that these neural systems are not simply associated with taking in positive or negative information, but rather with encoding and integrating social information that clashes with one's goals (here, forming positive impressions of ingroup members and negative impressions of outgroup members). Under such a model, dACC activation may signal a conflict between one's motivations and incoming, goal-inconsistent information, whereas LPFC activation may guide regulation efforts to resolve the conflict and encode such information (e.g. Egner, 2009; Braver, 2012; Shenhav *et al.*, 2014).

A related possibility is that dACC, MPFC, LPFC, TPJ and precuneus may encode prediction errors in social learning and decision-making (e.g. Suzuki *et al.*, 2012; Hackel *et al.*, 2015; Stanley, 2015). Negative information likely represents a prediction error when learning about ingroup (but not outgroup) members, and neural activation in these regions should increase to the extent that a prediction error is detected. However, positive information likely represents a prediction error when learning about outgroup members, but we failed to detect a significant relationship between activation in these regions and positive information-processing about outgroup members. One possibility, as we suggest earlier, is that the information presented in the current task is largely negatively skewed. One way to test whether prediction errors to positive information about outgroup members leads to improved outgroup impressions would be to examine contexts in which people learn largely positive information about social targets. Future research should refine the specific roles of neural systems involved in reducing biases in impression formation processes.

Second, these findings contribute to a growing body of research on the neuroscience of motivated cognition (see Hughes and Zaki, 2015 for a recent review). Nascent evidence suggests that motivated biases are associated with reduced activation in regions associated with executive control (e.g. LPFC, dACC/MPFC; Anderson *et al.*, 2004; Egner, 2009; Braver, 2012; Shenhav *et al.*, 2014), which suggests that such biases may be associated with effortless information-processing strategies. For example, a failure to appropriately consider undesirable information (as reflected by reduced LPFC, MPFC and dACC activation) is associated with a host of positively biased evaluations about oneself and motivationally relevant others (Krusemark *et al.*, 2008; Kato *et al.*, 2009; Beer and Hughes, 2010; Somerville *et al.*, 2010; Sharot *et al.*, 2011; Hughes and Beer, 2012a,b). Here, we demonstrate that the relatively effortless features of motivated cognition extend to impression formation. Specifically, 'reduced' activation in LPFC, dACC and other regions while learning negative information about ingroup members predicted 'more favorable' subsequent impressions about them, and activation in this system predicted a reduction of this ingroup bias.

This pattern of results is consistent with an effortless—rather than effortful—motivated information processing strategy (Baumeister and Newman, 1994; Sedikides and Green, 2000; Sanitioso and Wlodarski, 2004; Sedikides and Gregg, 2008). An effortful motivated information processing strategy involves actively suppressing unwanted information. For example, when people explicitly suppress encoding unwanted information, successful suppression is associated with increased LPFC activation (Benoit and Anderson, 2012). In contrast, the present findings suggest that people may sometimes passively avoid unwanted information, unless they engage LPFC, dACC, TPJ and additional regions to counteract this bias. Such findings suggest that people may heed motives by encoding information that threatens their desired conclusions in a relatively shallow manner. Tuning attention towards undesirable information may therefore help to reduce effortless biases in information processing.

The present findings dovetail with an emerging literature on the Bayesian brain hypothesis (Knill and Pouget, 2004). Under this model, rational observers make decisions by integrating their prior beliefs with new information through Bayes' rule. However, observers have adapted to integrate rationally within the confines of bounded resources (Jones and Love, 2011; Gershman *et al.*, 2015). These limitations give rise to a number of heuristics and biases in learning and decision-making. For example, an observer may deviate from ideal use of information by overweighing previously held beliefs (e.g. Achtziger *et al.*, 2014). This is consistent with the notion that observers often arrive at a conclusion they want to believe before they evaluate any evidence that might support or contradict such a conclusion (Kunda, 1990; Dunning, 2014). Participants in the current study might have likewise overweighed prior beliefs about ingroup members in the face of new countervailing evidence, and failed to integrate this new evidence into their judgment. To the extent that participants engaged greater activation of dACC/MPFC, LPFC, TPJ, Precuneus and additional regions while they encoded new evidence, the more likely they were to update their beliefs about a given ingroup member. This interpretation is consistent with findings from a number of social learning tasks that demonstrate a role for dACC, LPFC, TPJ, Precuneus and other regions in accumulating new evidence and integrating such evidence to guide future decisions (Behrens *et al.*, 2008; Stern *et al.*, 2010; Vilares *et al.*, 2012; d'Acremont *et al.*, 2013; O'Reilly *et al.*, 2013; Hackel *et al.*, 2015; Stanley, 2015; Zaki *et al.*, 2016).

Our findings also contribute to a large body of research on intergroup cognition. Neuroscientific work on intergroup processes largely focuses on defining groups along racial divides. In these interracial contexts, people generally attempt to suppress or control their negative race-based attitudes in order to appear unprejudiced (Devine, 1989; Dovidio *et al.*, 1997; Payne, 2001). However, other intergroup contexts differ from race in a number of ways. When groups are characterized by factors other than race such as competition, people often openly broadcast their beliefs about the superiority of their own group and/or their disregard for other groups (Tajfel, 1982; Brewer, 1999; Cikara and Van Bavel, 2014). Here, we demonstrate that activation in the very same systems associated with reducing racial biases, namely LPFC and dACC (e.g. Richeson *et al.*, 2003; Cunningham *et al.*, 2004; Amodio *et al.*, 2008), also reduces the expression of ingroup favoritism. These findings suggest that just as racial biases are expressed automatically in race-based intergroup contexts, ingroup favoritism may be expressed automatically in competitive intergroup contexts. One fruitful avenue for future research is to examine whether outgroup derogation in threatening or competitive outgroup contexts would also be associated with effortless bias. From this perspective, greater engagement in LPFC and dACC may be associated with reduced outgroup derogation.

Finally, the current findings suggest strategies to curb the consequences of motivation on cognition and behavior. A number of motivations and biases shape various stages of information processing (Hughes and Zaki, 2015), beginning with rapid initial perceptions of oneself, other people and the world (e.g. Balcetis and Dunning, 2006; Epley and Whitchurch, 2008; Van Bavel *et al.*, 2008; Ratner *et al.*, 2014) and culminating in real-world decisions and behaviors (e.g. Babcock and Loewenstein, 1997; Levine *et al.*, 2005; Moss-Racusin *et al.*, 2012). For example, physically attractive job candidates are evaluated more positively than less attractive candidates despite similar levels of competence (Bhanji and Beer, 2013), and male job candidates are offered higher starting salaries and more career mentoring than female job candidates with identical credentials (Moss-Racusin *et al.*, 2012). The current findings suggest that these and other effects may stem from fundamentally different encoding strategies that observers apply to information they encounter to align such information with their expectations and goals.

In this way, the current findings suggest fruitful avenues for future research on interventions designed to reduce conflict and inequality. Most interventions designed to reduce conflict and inequality increase intergroup contact and foster a sense of common identity between individuals. Although these interventions may promote positive interpersonal relations, they might also be limited in their ability to curtail downstream conflict and inequality (Dixon *et al.*, 2010). This is because similar information—for instance about the traits and behaviors of individuals—may be processed in fundamentally different ways as a function of the motivations of observers. The current work suggests the possibility that interventions that encourage effortful deliberation and deeper information-processing while encoding new information might curb the cascading consequences of motivation on cognition and reduce downstream inequality.

Acknowledgements

We thank Henry Tran, Jack Maris, and Omri Raccah for assistance during data collection, stimulus generation and task programming, Peter Mende-Siedlecki, Jamil Bhanji, Nicholas Camp, Takuya Sawaoka, Aneeta Rattan and Sylvia Morelli for helpful discussion and feedback, Stanford Center for Cognitive and Neurobiological Imaging for scanning support, and two anonymous reviewers for constructive suggestions.

Conflict of interest. None declared.

References

Achtziger, A., Alós-Ferrer, C., Hügelschäfer, S., Steinhauser, M. (2014). The neural basis of belief updating and rational decision making. *Social Cognitive and Affective Neuroscience*, **9**(1), 55–62.

Ambady, N., Bernieri, F.J., Richeson, J.A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In Zanna, M.P., editor. *Advances in Experimental Social Psychology*, Vol. **32**, pp. 201–272, San Diego, CA: Academic Press.

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.

Anderson, N.H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, **70**, 394.

Anderson, M.C., Ochsner, K.N., Kuhl, B., *et al.* (2004). Neural systems underlying the suppression of unwanted memories. *Science*, **303**, 232–5.

Aronson, E., Willerman, B., Floyd, J. (1966). The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science*, **4**, 1966, 227–8.

Asch, S.E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, **41**, 258.

Babcock, L., Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *The Journal of Economic Perspectives*, 109–26.

Balcetis, E., Dunning, D. (2006). See what you want to see: motivational influences on visual perception. *Journal of Personality and Social Psychology*, **91**(4),612.

Baron, S.G., Gobbini, M.I., Engell, A.D., Todorov, A. (2011). Amygdala and dorsomedial prefrontal cortex responses to appearance-based and behavior-based person impressions. *Social Cognitive and Affective Neuroscience*, **6**, 572–81.

Baumeister, R., Bratslavsky, E., Finkenauer, C., Vohs, K. (2001). Bad is stronger than good. *Review of General Psychology*, **5**, 323–70.

Baumeister, R.F., Newman, L.S. (1994). Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin*, **20**, 3–19.

Beer, J.S., Hughes, B.L. (2010). Neural systems of social comparison and the "above-average" effect. *Neuroimage*, **49**, 2671–9.

Beer, J.S., Chester, D.S., Hughes, B.L. (2013). Social threat and cognitive load magnify self-enhancement and attenuate self-deprecation. *Journal of Experimental Social Psychology*, **49**(4),706–11.

Behrens, T.E., Hunt, L.T., Woolrich, M.W., Rushworth, M.F. (2008). Associative learning of social value. *Nature*, **456**(7219),245–9.

Benoit, R.G., Anderson, M.C. (2012). Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron*, **76**(2),450–60.

Bhanji, J.P., Beer, J.S. (2013). Dissociable neural modulation underlying lasting first impressions, changing your mind for the better, and changing it for the worse. *The Journal of Neuroscience*, **33**, 9337–44.

Braver, T.S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, **16**, 106–13.

Brewer, M.B. (1999). The psychology of prejudice: ingroup love and outgroup hate?. *Journal of Social Issues*, **55**, 429–44.

Cikara, M., Van Bavel, J.J. (2014). The neuroscience of intergroup relations: an integrative review. *Perspectives on Psychological Science*, **9**(3),245–74.

Cloutier, J., Gabrieli, J.D.E., O'Young, D., Ambady, N. (2011b). An fMRI study of violations of social expectations: when people are not who we expect them to be. *NeuroImage*, **57**, 583–8.

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**, 806–13.

d'Acremont, M., Schultz, W., Bossaerts, P. (2013). The human brain encodes event frequencies while forming subjective beliefs. *The Journal of Neuroscience*, **33**(26),10887–97.

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

Dixon, J., Tropp, L.R., Durrheim, K., Tredoux, C. (2010). "Let Them Eat Harmony" Prejudice-Reduction Strategies and Attitudes of Historically Disadvantaged Groups. *Current Directions in Psychological Science*, **19**, 76–80.

Dovidio, J.F., Gaertner, S.L. (2010). Intergroup bias. In Fiske S. T., Gilbert D., & Lindzey G. (Eds.), *Handbook of Social Psychology*, 5th edn, Vol. **2**, pp. 1084–121, New York: Wiley.

Dovidio, J.F., Kawakami, K., Johnson, C., Johnson, B., Howard, A. (1997). On the nature of prejudice: automatic and controlled processes. *Journal of Experimental Social Psychology*, **33**, 510–40.

Dunning, D. (2014). Motivated cognition in self and social thought. In: Mikulincer, M., Shaver, P., Borgida, E., Bargh, J., editors. *APA Handbook of Personality and Social Psychology: Attitudes in social cognition*, Vol. **1**, pp. 777–803. Washington, DC: APA.

Egner, T. (2009). Prefrontal cortex and cognitive control: motivating functional hierarchies. *Nature Neuroscience*, **12**, 821–2.

Epley, N., Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin*, **34**, 1159–70.

Fiske, S.T. (1980). Attention and weight in person perception: the impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, **38**, 889–906.

Foddy, M., Platow, M.J., Yamagishi, T. (2009). Group-based trust in strangers the role of stereotypes and expectations. *Psychological Science*, **20**, 419–22.

Freeman, J.B., Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological review*, **118**(2), 247.

Freeman, J.B., Schiller, D., Rule, N.O., Ambady, N. (2010). The neural origins of superficial and individuated judgments about ingroup and outgroup members. *Human Brain Mapping*, **31**, 150–9.

Fuhrman, R.W., Bodenhausen, G.V., Lichtenstein, M. (1989). On the trait implications of social behaviors: kindness, intelligence, goodness, and normality ratings for 400 behavior statements. *Behavior Research Methods, Instruments, and Computers*, **21**, 587–97.

Gershman, S.J., Horvitz, E.J., Tenenbaum, J.B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, **349**(6245), 273–8.

Hackel, L.M., Doll, B.B., Amodio, D.M. (2015). Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience*, **18**, 1233–5.

Hamilton, D.L., Zanna, M.P. (1972). Differential weighting of favorable and unfavorable attributes in impressions of personality. *Journal of Experimental Research in Personality* **6**, 204–12.

Howard, J., Rothbart, M. (1980). Social categorization and memory for ingroup and outgroup behavior. *Journal of Personality and Social Psychology*, **58**, 301–8.

Hughes, B.L., Beer, J.S. (2012a). Orbitofrontal cortex and anterior cingulate cortex are modulated by motivated social cognition. *Cerebral Cortex*, **22**, 1372–81.

Hughes, B.L., Beer, J.S. (2012b). Medial orbitofrontal cortex is associated with shifting decision thresholds in self-serving cognition. *Neuroimage*, **61**(4), 889–98.

Hughes, B.L., Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, **19**, 62–4.

Jones, M., Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, **34**(4), 169–88.

Kato, J., Ide, H., Kabashima, I., Kadota, H., Takano, K., Kansaku, K. (2009). Neural correlates of attitude change following positive and negative advertisements. *Frontiers in Behavioral Neuroscience*, **3**, 1–13.

Knill, D.C., Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, **27**(12), 712–9.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, **63**(279), 190–4.

Krusemark, E.A., Keith Campbell, W., Clementz, B.A. (2008). Attributions, deception, and event related potentials: an investigation of the self-serving bias. *Psychophysiology*, **45**, 511–5.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin* **108**(3),480.

Lench, H.C., Ditto, P.H. (2008). Automatic optimism: Biased use of base rate information for positive and negative events. *Journal of Experimental Social Psychology* **44**(3),631–9.

Levine, M., Prosser, A., Evans, D., Reicher, S. (2005). Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shape helping behavior. *Personality and Social Psychology Bulletin*, **31**, 443–53.

Luhtanen, R., Crocker, J. (1992). A collective self-esteem scale: self-evaluation of one s social identity. *Personality and Social Psychology Bulletin*, **18**(3),302–18.

Ma, N., Vandekerckhove, M., Baetens, K., Van Overwalle, F., Seurinck, R., Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, **7**, 937–50.

Maass, A., Salvi, D., Arcuri, L., Semin, G.R. (1989). Language use in intergroup contexts: the linguistic intergroup bias. *Journal of Personality and Social Psychology*, **57**(6), 981.

Mende-Siedlecki, P., Cai, Y., Todorov, A. (2013a). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, **8**, 623–31.

Mende-Siedlecki, P., Baron, S.G., Todorov, A. (2013b). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *The Journal of Neuroscience*, **33**, 19406–15.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America*,, **109**(41),16474–9.

Mulder, M.J., Wagenmakers, E.J., Ratcliff, R., Boekel, W., Forstmann, B.U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience*, **32**(7), 2335–43.

O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E., Mars, R.B., Rushworth, M.F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(38), E3660–9.

Payne, K.B. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, **81**, 181–92.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences*, **11**(1),37–43.

Ratner, K.G., Dotsch, R., Wigboldus, D.H., van Knippenberg, A., Amodio, D.M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, **106**, 897.

Reeder, G.D., Brewer, M.B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, **86**, 61–79.

Richeson, J.A., Baird, A.A., Gordon, H.L., *et al.* (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, **6**, 1323–8.

Roy, M., Shohamy, D., Wager, T.D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, **16**(3),147–56.

Sanitioso, R.B., Wlodarski, R. (2004). In search of information that confirms a desired self-perception: motivated processing of social feedback and choice of social interactions. *Personality and Social Psychology Bulletin*, **30**(4),412–22.

Sedikides, C., Green, J.D. (2000). On the self-protective nature of inconsistency-negativity management: using the person memory paradigm to examine self-referent memory. *Journal of Personality and Social Psychology*, **79**, 906.

Sedikides, C., Gregg, A.P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science*, **3**, 102–16.

Schiller, D., Freeman, J.B., Mitchell, J.P., Uleman, J.S., Phelps, E.A. (2009). A neural mechanism of first impressions. *Nature Neuroscience*, **12**, 508–14.

Sharot, T., Korn, C.W., Dolan, R.J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, **14**, 1475–9.

Shenhav, A., Straccia, M.A., Cohen, J.D., Botvinick, M.M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Euroscience*, **17**, 1249–54.

Skowronski, J.J., Carlston, D. (1989). Negativity and extremity biases in impression formation: a review of explanations. *Psychological Bulletin*, **105**, 131–42.

Somerville, L.H., Kelley, W.M., Heatherton, T.F. (2010). Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cerebral Cortex*, **20**, 3005–13.

Stanley, D.A. (2015). Getting to know you: general and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience*, **11**, 525–36.

Stern, E.R., Gonzalez, R., Welsh, R.C., Taylor, S.F. (2010). Updating beliefs for a decision: neural correlates of uncertainty and underconfidence. *The Journal of Neuroscience*, **30**(23),8032–41.

Suzuki, S., Harasawa, N., Ueno, K., *et al.* (2012). Learning to simulate others' decisions. *Neuron*, **74**(6), 1125–37.

Tajfel, H., Turner, J.C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, **33**, 47.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, **33**, 1–39.

Taylor, S.E., Brown, J.D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, **103**(2),193.

Van Bavel, J.J., Packer, D.J., Cunningham, W.A. (2008). The neural substrates of in-group bias. *Psychological Science*, **19**, 1131–9.

Vilares, I., Howard, J.D., Fernandes, H.L., Gottfried, J.A., Kording, K.P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology*, **22**(18), 1641–8.

West, R.F., Meserve, R.J., Stanovich, K.E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, **103**(3), 506.

Zaki, J., Kallman, S., Wimmer, G.E., Ochsner, K., Shohamy, D. (2016). Social cognition as reinforcement learning: feedback modulates emotion inference. *Journal of Cognitive Neuroscience*, **28**, 1270–82.

Zaki, J., Hennigan, K., Weber, J., Ochsner, K.N. (2010). Social cognitive conflict resolution: contributions of domain-general and domain-specific neural systems. *The Journal of Neuroscience*, **30**(25), 8481–8.

Zaki, J. (2013). Cue integration a common framework for social cognition and physical perception. *Perspectives on Psychological Science*, **8**(3), 296–312.