Motivation alters impression formation and related neural systems

Brent L. Hughes,1,2 Jamil Zaki,1 and Nalini Ambady1,†

1Department of Psychology, University of California, Riverside, CA 92521, USA and 2Department 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.
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 typically 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., 2011b; 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
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).

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 (2.8-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 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 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 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 individually 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 LFPC and dACC for visualization purposes alone, as these regions are (i) consistently activated across all of our reported analyses and (ii)
after’ the first impressions were made; all t(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 overweight 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.

<table>
<thead>
<tr>
<th>Region of Activation</th>
<th>BA</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>T-stat</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DACC</td>
<td>24</td>
<td>0</td>
<td>18</td>
<td>38</td>
<td>5.57</td>
<td>167</td>
</tr>
<tr>
<td>TPJ/IPL</td>
<td>39/40</td>
<td>48</td>
<td>−46</td>
<td>22</td>
<td>4.76</td>
<td>172</td>
</tr>
<tr>
<td>Insula</td>
<td>−56</td>
<td>10</td>
<td>6</td>
<td>4.69</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>LPFC</td>
<td>36</td>
<td>14</td>
<td>0</td>
<td>4.21</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Precuneus</td>
<td>7</td>
<td>−10</td>
<td>10</td>
<td>28</td>
<td>3.61</td>
<td>26</td>
</tr>
</tbody>
</table>

**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).

**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 dIPFC, 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
negative information (e.g. Bhanji and Beer, 2013; Mendes-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 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.
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 countering this bias engages greater activation in neural systems associated with impression updating.

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

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

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

**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.
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, TPJ, insula, precuneus and other regions (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

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

<table>
<thead>
<tr>
<th>Region of Activation</th>
<th>BA</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>T-stat</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingroup Negative Encoding &gt; Outgroup Negative Encoding Contrast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPFC</td>
<td>44/45</td>
<td>50</td>
<td>28</td>
<td>28</td>
<td>4.50</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>-44</td>
<td>34</td>
<td>28</td>
<td>3.95</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>TPJ/IPL</td>
<td>39/40</td>
<td>50</td>
<td>-60</td>
<td>24</td>
<td>3.61</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>-36</td>
<td>-58</td>
<td>22</td>
<td>3.91</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>Insula</td>
<td>-32</td>
<td>18</td>
<td>-11</td>
<td>3.50</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>Parahippocampal cortex</td>
<td>32</td>
<td>-22</td>
<td>-14</td>
<td>3.50</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>Temporal cortex</td>
<td>-45</td>
<td>-12</td>
<td>-14</td>
<td>3.46</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Ingroup Negative Encoding &gt; Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPFC</td>
<td>44/45</td>
<td>52</td>
<td>20</td>
<td>30</td>
<td>4.14</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>50</td>
<td>-4</td>
<td>4.23</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-40</td>
<td>24</td>
<td>38</td>
<td>3.41</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>TPJ/IPL</td>
<td>39/40</td>
<td>60</td>
<td>-54</td>
<td>14</td>
<td>4.08</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>-60</td>
<td>-52</td>
<td>36</td>
<td>4.00</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-45</td>
<td>-58</td>
<td>22</td>
<td>3.70</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Insula</td>
<td>-45</td>
<td>14</td>
<td>-4</td>
<td>4.08</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>20</td>
<td>-8</td>
<td>3.47</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>DMPFC</td>
<td>8/9</td>
<td>8</td>
<td>22</td>
<td>54</td>
<td>3.57</td>
<td>45</td>
</tr>
<tr>
<td>DACC</td>
<td>24</td>
<td>2</td>
<td>24</td>
<td>32</td>
<td>3.21</td>
<td>23</td>
</tr>
</tbody>
</table>

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
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 intercultural 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; Oake 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 multiple 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 intergroup 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 Sawaoaka, 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


