Running Title: Parametric modulation of FCE

Disconfirmation Modulates the Neural Correlates of the False Consensus Effect:

A Parametric Modulation Approach

B. Locke Welborn[1]*

Matthew D. Lieberman[2]

[1]SAGE Center for the Study of the Mind, University of California, Santa Barbara

[2]University of California, Los Angeles

Correspondence should be addressed to:

Benjamin Locke Welborn
Sage Center for the Study of the Mind
2213 Psychology, UCSB
Santa Barbara, CA 93106
Phone: 203.710.0030
Email: locke.welborn@sagecenter.ucsb.edu

Abstract

The False Consensus Effect (FCE) – the tendency to (erroneously) project our attitudes and opinions onto others – is an enduring bias in social reasoning with important societal implications. In this fMRI investigation, we examine the neural correlates of within-subject variation in consensus bias on a variety of social and political issues. Bias demonstrated a strong association with activity in brain regions implicated in self-related cognition, mentalizing, and valuation. Importantly, however, recruitment of these regions predicted consensus bias only in the presence of social disconfirmation, in the form of feedback discrepant with participants' own attitudes. These results suggest that the psychological and neural mechanisms underlying the tendency to project attitudes onto others are crucially moderated by motivational factors, including the desire to affirm the normativity of one's own position. This research complements social psychological theorizing about the factors contributing to the FCE, and further emphasizes the role of motivated cognition in social reasoning.

INRODUCTION

The false consensus effect (FCE) – the tendency to (erroneously) presume that others share our attitudes, opinions, and beliefs – is one of the most pervasive and recalcitrant biases in human social reasoning (Ross, Green, & House, 1977; Marks & Miller, 1987). We persist in projecting our own minds on to others, even when we are made aware of the possibility of such consensus estimation bias and even when we acknowledge bias in others' judgments (Krueger & Clement, 1994; Pronin, Lin, & Ross, 2002; Pronin, Gilovich, & Ross, 2004). Our own opinions

simply seem to 'count' for more, though our goal in evaluating group consensus is to objectively consider the thoughts and feelings of individuals who may be very different from us. Numerous psychological theories have been advanced to explain this phenomenon, but no single hypothesis seems likely to account for all cases of bias. Rather, some combination of motivated social reasoning, selective exposure to similar others, and the chronic accessibility of the self may activate consensus bias in concrete situations and sustain it against attempts at control (Sherman, Presson, Chassin, et al., 1983; Clement & Krueger, 2002). However, it is unclear how these putative mechanisms contribute differentially to consensus bias, and under what kinds of circumstances.

In particular, the role of motivated cognition in driving consensus bias has been a point of contention. Several studies have found consensus bias to be stronger when individuals have a strong need to justify their actions (Sherman, Presson, Chassin, et al., 1983; Wolfson, 2000), have a vested interest in social consensus (Crano, 1983), or when the self is placed under threat (Sherman, Presson, & Chassin, 1984; Morrison & Matthes, 2011). And suggestively, the less common an attitude actually is in the population, the *more* likely its proponents are to over-estimate its prevalence (Krueger & Clement, 1997; Mullen & Hu, 1988; Mullen & Smith, 1990). However, while these results are consistent with the notion that individuals (especially those in the minority) are motivated to assert and defend the normativity of their positions, it does not rule out alternative (non-motivational) explanations. Overall, the role of motivated processes in consensus bias remains unclear, and alternative mechanistic explanations for bias have proven difficult to disambiguate experimentally.

Given this impasse, the tools of functional neuroimaging may provide a useful means for testing hypotheses regarding the factors that contribute to consensus bias and for examining their

interactions in real-time. If motivated cognition plays a role in consensus bias, we would expect its influence to be emphasized under conditions of social threat, when motivation to defend the self is high (Hughes & Beer, 2010; Morrison & Matthes 2011). For instance, finding out that another person's attitude is discrepant with one's own (i.e. *disconfirmation*) right before making a consensus judgment might enlist motivated processes to a greater degree. These processes may then sustain bias in the face of discrepant social feedback, effectively discounting the attitudes of those who disagree with us as uninformative or irrelevant to assessing group consensus. In so doing, we may be able to maintain the conviction that our attitudes and beliefs are in the majority, and buffer ourselves affectively from the consequences of minority status.

Interpretation of neuroimaging data associated with consensus bias is aided considerably by the burgeoning literature on the neural correlates of motivated cognition (Beer & Hughes, 2010; for review, see Hughes & Zaki, 2015). Motivated cognition depends upon regions involved in computations of subjective value (i.e. social and non-social rewards and costs), self-related processes, and mentalizing (i.e. thinking about others mental states). Each of these sub-components has been the subject of extensive research (c.f. Bartra, McGuire, & Kable, 2013 regarding subjective value, Murray, Schaer, & Debané, 2012; Northoff et al., 2006 for self-related cognition, and Van Overwalle, 2009; Molenberghs et al., 2016 for reviews of work on mentalizing). While these processes all involve regions collectively associated with social cognition (Lieberman, 2010) and the default mode (Raichle, 2015), they are also empirically dissociable, as observed in the aforementioned reviews and in automated meta-analysis (for example, through www.neurosynth.org).

While mentalizing regions have featured less prominently in contemporary neuroimaging work on motivated cognition than regions associated with valuation and self-related cognition,

we believe they are especially pertinent to the representation (and misrepresentation) of others' attitudes. Indeed, mentalizing regions seem in particular to be implicated in thinking about the relationship between our own attitudes and those of others, for example, when we form intentions to share persuasive messages (Falk et al., 2013).

The hypothesis that motivated processes influence consensus estimation yields clear predictions regarding the involvement of self, valuation, and mentalizing regions in consensus bias. To the extent that social contexts challenge or threaten the self, neural mechanisms involved in motivated cognition (especially those implicated in and computations of subjective value, self-related cognition, and mentalizing) should show altered associations with the magnitude of exhibited consensus bias. In particular, motivational accounts of the false consensus effect emphasize the importance of maintaining the belief that our own attitudes and behaviors are reasonable and normative (Sherman, Presson, & Chassin, 1984). Thus, we might predict that the engagement of psychological and neural mechanisms supporting bias should be modulated by the presence of motivations to defend the 'majority status' of our own beliefs against possible challenges. When we have reason to believe that others may disagree with us, we may engage additional psychological processes when considering their mental states and comparing them to our own. Indeed, discrepant feedback places us in something of a conundrum or cross-roads as social thinkers: on the one hand, social disagreement provides evidence that our attitudes and beliefs may not be as common or pervasive as we previously thought – on the other hand, it could spur us to defensively reassert our majority status, 'doubling-down' on projective bias. In other words, discrepant feedback motivates us to determine whether it is *us* or the *other individual* who is out of step with the consensus view.

Moreover, in Welborn, Gunter, Vezich, & Lieberman (2017), between-subjects variation in observed consensus bias was associated with the recruitment of reward regions such as the nucleus accumbens (NAcc) and the ventromedial prefrontal cortex. That is, individuals who exhibited greater activation in these regions, on average, tended to show greater bias in their consensus estimates. Conversely, a region implicated in emotion regulation (the right ventrolateral prefrontal cortex, or RVLPFC) was inversely related to observed consensus bias across subjects. These results point to the possibility that motivated processes may contribute to bias, but are limited to comparisons across individuals. We do not yet know, crucially, whether variation in bias across attitude items is associated with differential recruitment of neural reward circuitry, within-subjects.

With these considerations in mind, we sought to characterize the neural correlates of consensus bias (or the FCE) in an undergraduate sample while they estimated the attitudes of an ordinary university student on a variety of different social issues. Because of our interest in the contextual factors shaping consensus estimation (in particular, the availability of social feedback regarding others' positions) we presented participants with three information conditions. In the first condition (No Information), participants merely provided their consensus estimates without any outside influence. In two comparison conditions, participants provided consensus estimates after learning that one of their peers either had a similar attitude (Confirmation) or a discrepant attitude (Disconfirmation). In evaluating the neural correlates of consensus bias, we focused on the technique of parametric modulations, because we sought to identify regions in which activity co-varied with the amount of bias exhibited on a trial-by-trial (or issue-by-issue) basis. On the basis of social psychological research on the FCE and the social neuroscience literature on motivated reasoning, we were especially interested in evaluating: 1) whether activation of self-

related (principally medial prefrontal cortex (MPFC, BA10) and precuneus), mentalizing (dorsomedial prefrontal cortex (DMPFC, BA8/9) and bilateral temporo-parietal junction (TPJ)), and value (ventromedial prefrontal cortex (VMPFC, medial BA11) and ventral striatum/nucleus accumbens) regions would be linearly associated with the magnitude of consensus bias observed for each attitude, and 2) whether the context of judgment (No Information / Confirmation / Disconfirmation) would moderate the neural responses in these regions.

If motivated processes play an important role in consensus bias, they should be associated with differential neural correlates across the experimental conditions manipulating the availability and nature relevant social information. Conversely, identical neural correlates of bias in the presence and absence of disconfirmation would not provide distinctive evidence for an account that prioritizes motivated cognition. In particular, if motivated processes increase consensus bias, we predict that the above-mentioned regions (implicated in mentalizing, valuation, and self-related cognition) should show strong, positive associations with bias in the Disconfirmation condition, and weaker associations with bias in the other conditions. Similar neural correlates across conditions, or differential neural correlates in regions not associated with motivated cognition in prior literature, would fail to support the importance of motivation in shaping consensus bias. Instead, such results might provide evidence in favor of alternative processes sustaining bias. For example, strong positive associations between bias and fronto-parietal activation might suggest the involvement of selective attention processes (Corbetta & Shulman, 2002), while engagement of anterior temporal lobe structures might suggest biased retrieval of social information relevant to attitudes (Wang et al., 2017).

Data and results from this sample of participants have been presented elsewhere (Welborn, Gunter, Vezich, & Lieberman, 2017), focusing on between-subjects differences in the

activation of areas underlying consensus bias. Here, rather than considering individual differences, we focus instead on common processes that are associated with within-subjects variation in consensus bias across issues (or attitude items). Some of the regions queried in the present study overlap with those interrogated in Welborn, Gunter, Vezich, & Lieberman (2017), insofar as the valuation regions include the NAcc and portions of the VMPFC (see ROI definition in *Methods* below), but the analysis performed is independent. In the *Results* section below, prior to discussing parametric modulation analysis, we address several aspects of the behavioral data that are relevant to parametric modulation analyses: 1) differences in reaction times across conditions. 2) correlations between reaction times and bias, and 3) the magnitude of variation in bias scores across conditions.

METHODS

The methods employed in the conduct of this research have previously been described in Welborn, Gunter, Vezich, & Lieberman (2017). They are reproduced here for convenience, with minor changes to explain the models used in assessing parametric modulation of neural activity by consensus bias.

Participants

Twenty-nine participants (17 female) were recruited by email and Internet solicitations from the psychology research subject pool at UCLA. All participants had been enrolled as undergraduate students at UCLA for at least two quarters, and none had taken an introductory course in social psychology (in order to preclude familiarity with the false consensus effect). Participants were judged ineligible if they did not differ from our estimate of the mean UCLA undergraduate attitude on a sufficient number of items. All participants were compensated $40 for their contribution to this research or received course credit. Participants provided written

informed consent approved by the UCLA Institutional Review Board. One participant's data are not included in these analyses due to partial acquisition failure (final n=28).

Attitude Item Selection

Attitude items were selected from a larger set of 155 social, political and personal issues (e.g. abortion rights, gay marriage, daily flossing, making out on a first date) that had previously been tested with an online sample of 178 UCLA undergraduates. Participants in this online sample indicated their attitudes towards each issue using a numeric scale ranging from 0 to 100 in integer increments (with anchors 0 – Complete Opposition, 25 – Moderate Opposition, 50 – Neutrality, 75 – Moderate Support, and 100 – Complete Support). These responses provided a reasonable estimate of the mean UCLA undergraduate attitude on each of the 155 issues, and these values were used to determine error of estimation for the scanner task described below.

Prior to scanning, prospective participants in the present study indicated their own attitudes on each of the 155 issues, and were eligible to participate only if their responses differed from our estimate of the UCLA undergraduate population mean by at least 15 points on at least 90 items. If participants did not differ in their attitudes from the group mean for the items used, it would not possible to disambiguate projection from accurate consensus estimation on a trial-by-trial level. As this was a major objective of the study, we felt it was necessary to impose such an inclusion criterion in order to provide a sufficient number of viable trials for the scanner task. The idiosyncrasies of participants' attitudes on the stimulus issues resulted in the selection of a unique set of attitude items for each individual, on each of which they differed from the UCLA undergraduate mean by at least 15 points. These items were randomly and equivalently divided amongst the Confirmation, Disconfirmation, and No Information conditions. Across

participants, this procedure resulted in an average of 99 trials total, or 33 per Consensus Estimation condition.

Consensus Estimation Task:

While undergoing functional magnetic resonance imaging (fMRI), participants estimated the attitude of the ordinary UCLA student on each of the ideographically-selected attitude items (see above). During the 'No Information' condition, participants were simply asked to provide their best possible estimate of the attitude that an ordinary UCLA student would have on the given issue. In order to do this, they used an on-screen scale identical to that used during item selection (as described above) except that the values represented the attitude that the ordinary UCLA student would have, rather than the participant's own attitude.

In the 'Confirmation' and 'Disconfirmation' conditions, participants were provided with on-screen information ostensibly reflecting the attitudes of other UCLA undergraduates. Participants were told that, on each trial, the attitude of a different UCLA student from our larger Internet sample would be presented, and that they could use (or disregard) this information in making their consensus estimates. While this sample actually existed, and was used to determine the true norms for each attitude item as described above, participants actually received false information designed to either Confirm or Disconfirm the presupposition that their own attitudes would be representative of the UCLA undergraduate population as a whole. In the Confirmation condition participants were provided with an attitude that differed from their own by at most 5 points (in either direction). As all attitude items were pre-selected so that participants attitudes were at least 15 points different from the mean, this ensured that the sample attitudes presented in the Confirmation were closer to the participant's own attitude than to the mean UCLA undergraduate attitude. In the Disconfirmation condition participants were provided with a

sample attitude that differed from the actual mean UCLA undergraduate attitude by at most 5 points (in either direction), so that this sample attitude was invariably closer to the actual mean than to the participant's own attitude. In both Confirmation and Disconfirmation conditions, deviations from the participant's own attitude and the mean UCLA undergraduate attitude were selected from a uniform random distribution so as to ensure that the presented attitude fell within the desired range.

On each trial (see Figure 1), the sample information (ostensibly reflecting the attitude of a single UCLA undergraduate) was presented numerically above the appropriate portion of the scale, with a line denoting the precise location corresponding to the other student's attitude. After the scale (and if applicable, sample information) had appeared on-screen, participants had 10 seconds within which to make their response. Trials were not explicitly separated into feedback and response phases, and sample information remained on-screen until participants had confirmed their response. Trial presentation was self-paced, with a jitter duration commencing immediately after participants' responses were registered. Inter-trial jitter was selected from an exponential random distribution with a range of 4-9s and a mean value of 5 seconds.

Non-social color-judgment trials were also included as a basic perceptual-motor control condition. On these trials, participants were asked to judge the color of an on-screen square that varied continuously from completely red to completely blue. Participants were instructed to treat the mid-point value of '50' as indicating that the square appeared to them completely purple, and neither more blue nor more red in hue. If the square appeared more red than blue, participants were to select values greater than 50, with 100 indicated that they perceived the square to be completely red. If the square appeared more blue than red, participants were to select values less than 50 with 0 indicated that the square completely blue. Participants were instructed explicitly

to provide *their own* judgment regarding the color of the square, and to ignore how others might perceive it. Thirty control trials were included in the task for each participant, intermixed with consensus estimation trials. Trial order was pseudo-randomized such that no condition repeated more than twice sequentially and conditions were represented equally over two functional runs.

fMRI data acquisition

All imaging data was acquired using a 3.0-Tesla Siemens Trio scanner at the Ahmanson-Lovelace Brain Mapping Center at UCLA. Across 2 functional runs, approximately 650 T2*-weighted echo-planar images were acquired during completion of experimental tasks described above (slice thickness=3mm, gap=1mm, 36 slices, TR=2000ms, TE=25ms, flip angle=90°, matrix=64x64, field of view=200mm). An oblique slice angle was used in order to minimize signal drop-out in ventral medial portions of the brain. In addition, a T2-weighted, matched-bandwith anatomical scan was acquired for each participant (TR=5000ms, TE=34ms, flip angle=90°, matrix=128x128; otherwise identical to EPIs). Lastly, we acquired a T1-weighted magnetically-prepared rapid acquisition gradient echo anatomical image (slice thickness=1mm, 176 slices, TR=2530ms, TE=3.31ms, flip angle=7°, matrix=256x256, field of view=256mm).

fMRI Data Preprocessing and Analysis

<u>Preprocessing:</u>

Functional data were analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Within each functional run, image volumes were corrected for slice acquisition timing, realigned to correct for head motion, segmented by tissue type, and normalized into standard MNI stereotactic space (resampled at 3x3x3mm). Finally, images were smoothed with an 8mm Gaussian kernel, FWHM.

<u>fMRI analytic paradigm:</u>

General linear models were defined for each participant, in which trials were modeled with separate functions corresponding to 1) the initial presentation of the trial and 2) a fixed epoch corresponding to the final 2.5 seconds preceding (and including) the participants' final response. The initial portion of the trial differs significantly between conditions, with the Confirmation and Disconfirmation conditions, but not the No Information condition, including on-screen information regarding the attitudes of another UCLA undergraduate. As parameter estimates from this portion of the trial are not directly comparable across conditions, the initial portion of each trial was therefore modeled as a parameter of no interest in the GLM. Parametric modulation analyses were conducted on parameter estimates corresponding to the final period of each trial (i.e. the last 2.5 seconds before participant response), which we believe better corresponds to the period of participants' decision-making and response selection. Both stimulus presentation and response selection were convolved with the canonical (double-gamma) hemodynamic response function.

The first (condition-agnostic) model collapsed across information conditions, in order to assess associations between hemodynamic activity and bias across all consensus estimation trials. In this model, two regressors of interest were modeled corresponding to the response period of consensus estimation trials (including Confirmation, Disconfirmation, and No Information) and of Control trials. One additional parameter of interest was included to model parametric modulation of response-period hemodynamic activity (irrespective of information condition) by the observed magnitude of consensus bias on each trial. In a second (condition-specific) model, four regressors of interest were modeled to the response period of the Confirmation, Disconfirmation, No Information, and Control conditions. In addition, three additional parameters of interest were included to model parametric modulation within each

condition of response-period hemodynamic activity by the observed magnitude of consensus bias across trials. Both models controlled for 18 motion parameters (3 translations and rotations, as well as their squares and first-order derivatives), and a junk regressor for acquisitions on which either translation exceeded 2mm or rotation exceeded 2 degrees in any direction. The time series was high-pass filtered using a cutoff period of 128s and serial autocorrelations were modeled as an autoregressive AR(1) process.

Consensus bias was computed on a trial-by-trial basis as the error of estimation of a participant's consensus estimate regarding the attitude item (relative to the true mean of our larger, 197 person sample) *in the direction of the participant's own attitude on the attitude item* (acquired several days before the scan). That is, consensus bias was operationalized as follows:

$$bias = \begin{cases} consensus\ estimate - true\ sample\ mean, if\ own\ attitude > true\ sample\ mean \\ true\ sample\ mean - consensus\ estimate, if\ own\ attitude < true\ sample\ mean \end{cases}$$

Consensus bias thus equals the (positive) magnitude of overestimation for items about which participant expresses above-average support, and the (positive) magnitude of underestimation for items about which the participant expresses below-average support. Bias values were also capped by the participant's own attitude; that is participants could not have a bias score greater than the difference between their own attitude and the sample mean. Capping bias values at the extremity (absolute value(own attitude – true sample mean)) of participants own attitudes ensures that participants cannot have strong bias scores when their estimates are more extreme than their own attitudes.[1] The consensus bias metric used is thus positive when participants overestimate support for their own attitudinal positions in the UCLA undergraduate population, negative when they underestimate support for their own attitudinal positions in the

[1] E.g. if own attitude = 60, sample mean = 50, and consensus estimate = 80, participants' bias is capped at 10; it would be unreasonable to attribute all 30 scale points of error to bias *towards one's own attitude* when a portion of that error move the estimate *away* from one's own position.

undergraduate population, and 0 if their estimate is accurate. Because this bias metric is sensitive to participants' actual over-estimation of support for their own attitudes, we believe it is an effective operationalization of consensus bias for the purposes of imaging research. It is conceptually similar to the 'truly false consensus effect' developed by Krueger and Clement (Krueger & Clement, 1994). Using these bias scores, parametric modulation analyses were conducted to identify regions in which hemodynamic activity co-varied with participants' bias on a trial-by-trial basis.

Individual-level statistics were aggregated for group-level comparisons and evaluated with a mixed-effects model. For whole-brain analyses, correction for multiple comparisons was implemented based upon Gaussian Random Field theory, to yield cluster FWE of $p<0.05$ based upon an initial (voxel-wise) cluster-formation threshold of $p<0.005$.

Region of Interest (ROI) Analysis:

Region-of-interest (ROI) analyses were conducted to directly assess the recruitment of self, mentalizing, and value regions in consensus bias (see Figure 2A). ROIs for these comparisons were derived from www.neurosynth.org reverse-inference using the terms 'self', 'mentalizing', and 'value', thresholded at a t-value of 5 and resliced into 3x3x3 MNI space. ROIs were further limited to clusters of greater than 20 contiguous voxels and constrained to be exclusive (i.e., non-overlapping). The 'self' ROI was constrained to medial prefrontal cortex and precuneus, while the mentalizing ROI included clusters in both DMPFC as well as bilateral temporo-parietal junction. The valuation ROI comprises clusters in both VMPFC as well as ventral striatum/nucleus accumbens. Analysis of these ROIs is meant to directly test the hypothesis that regions relative to motivated cognitive processes will show differential neural correlates across information conditions (e.g., specifically in the presence of social

disconfirmation). Parameter estimates from the models described above were extracted from all ROIs using MarsBaR (Brett, Anton, Valabregue, & Poline, 2002) for statistical comparisons. Statistical tests reported are two-tailed.

RESULTS

The behavioral results from this experiment have previously been reported and are summarized briefly here for convenience. Interested readers are encouraged to consult Welborn, Gunter, Vezich, & Lieberman (2017) for further details.

Consistent with the extensive behavioral literature on the false consensus effect, consensus bias scores were significantly greater than zero both overall and for each information condition individually (M_{all}=12.17, t(27)=15.265, p<0.001; M_{Con}=19.07, t(27)=18.604, p<0.001; M_{NoI}=10.32, t(27)=9.950, p<0.001; M_{Dis}=8.27, t(27)=10.445, p<0.001). Repeated-measures analysis of variance revealed a substantial effect of information condition (Confirmation, Disconfirmation, or No Information) on participants' exhibited bias (F(2,54)=80.58, p<0.001). Participants showed greater bias in the Confirmation condition than the No Information condition (M_{Con}=19.07versus M_{NoI}=10.32, t(27)=9.095, p<0.001). Participants also showed significantly less bias in the Disconfirmation condition than in either the No Information condition (M_{Dis}=8.27 versus M_{NoI}=10.31, t(27)=-2.279, p=0.031) or the Confirmation condition (M_{Dis}=8.27 versus M_{Con}=19.07, t(27)=-11.509, p<0.001). These results suggest that participants are strongly susceptible to bias, over-estimating support for their own attitudinal positions by between 12 points out of a 100-point scale, on average. Participants are also sensitive to presentation of social information regarding the attitudes of their peers, and adjust their consensus estimates in light of this feedback. However, we emphasize that mean bias was significantly greater than zero for *all conditions*.

The presentation of sample information also affected participants' reaction times ($F(2,54)=5.137$, $p=0.007$). Predictably, both the Confirmation and Disconfirmation conditions resulted in longer reaction times than the No Information condition ($M_{Con}=4.54$ versus $M_{NoI}=4.32$, $t(27)=3.077$, $p=0.005$; $M_{Dis}=4.52$ versus $M_{NoI}=4.32$, $t(27)=2.367$, $p=0.025$). However, the Confirmation and Disconfirmation conditions did not differ in reaction time ($M_{Con}=4.54$ versus $M_{Dis}=4.52$, $t(27)=0.274$, $p=0.786$). Overall, correlation between mean consensus bias and mean reaction time was not significant, averaging across all conditions ($r=-0.344$, $p=0.073$). Mean bias in the Confirmation condition was inversely correlated with mean reaction time to Confirmation trials ($r=-0.399$, $p=0.035$), but this relationship did not hold for the Disconfirmation or No Information conditions. Moreover, mean intra-subject variation in bias was not correlated with mean intra-subject reaction times for any condition (Confirmation: $r(26)=-0.14$, $p=0.48$; No Information: $r(26)=-0.01$, $p=0.96$; Disconfirmation: $r(26)=-0.13$, $p=0.50$). These results suggest that differences in reaction time are unlikely to account for differences in consensus bias, either between-subjects or within-subjects. Moreover, comparable reaction times across conditions suggest that parametric modulation results will not have drastically different meaning or power across conditions.

Given the intent to assess parametric modulations with observed consensus bias, it is also important to rule out possible confounds connected to within-subject variance in consensus bias scores. Importantly, if there is a restriction of range of consensus bias scores in some conditions (e.g. if participants show high bias consistently for all trials in the Confirmation condition), it might be difficult to evaluate parametric modulations. The variation of consensus bias scores was highest in the No Information condition (mean intra-subject SD=18.00), intermediate in the No Information condition (mean intra-subject SD=16.93), and lowest in the Disconfirmation

condition (mean intra-subject SD=14.29). Variation was significantly higher in the No Information condition relative to the Disconfirmation condition ($t(27)=4.95$, $p<0.001$) and in the Confirmation condition relative to the Disconfirmation condition ($t(27)=3.98$, $p<0.001$), but did not significantly different between the No Information and the Confirmation conditions ($t(27)=1.59$, $p=0.12$). This suggests that participants were moderating their consensus estimates in the Disconfirmation condition, as expected, resulting in fewer extremely biased responses. Because of the lower predictor variance associated with consensus bias scores in the Disconfirmation condition, we may have reduced power to detect parametric modulations for this condition.

Overall, these results suggest that participants integrated the sample information into their consensus estimates as expected, showing greater bias in the presence of social confirmation and reduced bias in the presence of disconfirmation. Weak correlations between bias and reaction times suggest that this factor does not represent a serious confound for parametric modulation analyses that follow. Roughly comparable variance in bias scores across conditions indicate that parametric modulation analyses are appropriate, with the caveat that we may have reduced power to detect effects in the Disconfirmation condition.

Parametric modulation in ROIs

Parametric modulation analyses were conducted to determine whether hemodynamic activity in the regions-of-interest (ROIs) co-varied with actually observed bias on a trial-by-trial (issue-by-issue) basis. Mean estimates for the parametric modulation of the bias are shown in Figure 2B for each ROI and condition of interest.

Hemodynamic activity demonstrated positive, linear parametric modulation by observed consensus bias in the Disconfirmation condition for the self ($t(27)=3.958$ $p=0.0005$), mentalizing

(t(27)=2.765, p=0.0101), and value (t(27)=3.732, p=0.0009) ROIs. In marked contrast, these same regions did not exhibit parametric modulation of activity by bias in the Confirmation and No Information conditions (Confirmation: $t_{self}(27)$=0.-0.0268, p=0.980; $t_{mentalizing}(27)$=-1.585, p=0.124; $t_{value}(27)$=0.369, p=0.715; No Information: $t_{self}(27)$=-0.375, p=0.711; $t_{mentalizing}(27)$=-0.760, p=0.454; $t_{value}(27)$=-0.116, p=0.908). Moreover, the magnitude of parametric modulation by bias was greater in the Disconfirmation condition than the Confirmation and No Information conditions for all three ROIs: self (t (Disconfirmation > Confirmation t(27)=2.614, p=0.0145; Disconfirmation > No Information t(27)=3.427, p=0.00197), mentalizing (Disconfirmation > Confirmation t(27)=2.787, p=0.0096; Disconfirmation > No Information t(27)=2.756, p=0.0104), and value (Disconfirmation > Confirmation t(27)=2.0687, p=0.0482; Disconfirmation > No Information t(27)=2.860, p=0.0081).

These results indicate that the recruitment of regions involved in self-related cognition, mentalizing, and computations of value was strongly influenced by the social/informational context in which judgments of consensus were made. In the Disconfirmation condition, but not the other conditions tested, consensus bias was associated with the level of activity in each of these regions. Thus, when faced with social disconfirmation in the form of discrepant feedback regarding another person's attitude, the recruitment of regions implicated in motivated cognition predicted persistence in biased consensus estimates.

Parametric modulation in whole-brain analyses

The ROIs analyzed above were selected in order to directly test whether or not regions implicated in motivated processes would show differential associations with bias under social disconfirmation. Whole-brain analysis were conducted both to clarify the spatial localization of regions associated with biased consensus estimation, as well as to identify areas outside the *a*

priori ROIs that might show similar effects. Across all trials (i.e. ignoring condition), greater trial-wise consensus bias was associated with increased activity in the medial prefrontal cortex and ventromedial prefrontal cortex (MPFC, BA10; VMPFC, BA11; see Table 1 and Figure 3A). However, when conditions were analyzed separately (see below), it became evident that this effect is driven by and ultimately specific to the Disconfirmation condition.

In the Disconfirmation condition, large clusters within self-related, mentalizing, and valuation regions (including MPFC, VMPFC, precuneus, left temporo-parietal junction (LTPJ), and left temporal pole) demonstrated positive associations between activity and consensus bias, with additional clusters identified in the right amygdala, right caudate nucleus, and the thalamus (see Table 1 and Figure 3B). Thus, when participants' belief in the normativity of their attitudes was directly challenged by feedback from a fellow undergraduate, broad recruitment in these regions was associated with the persistence of bias. In marked contrast, in the Confirmation and No Information conditions, no significant clusters were found to exhibit parametric modulation with observed consensus bias, either positively or negatively. Moreover, when analyzing all non-Disconfirmation trials (i.e. Confirmation and No Information trials taken together) parametric modulation by consensus bias was only observed in a limited cluster within the superior parietal lobule.

In order to explicitly test whether the parametric engagement of self-related and mentalizing regions in consensus bias was specific to a state in which the predominance of one's own attitudes had been challenged (i.e. the Disconfirmation condition), the estimates of parametric bias modulation were compared between Disconfirmation and non-Disconfirmation trials. When compared to all other consensus estimation trials together (Confirmation and No

Information trials), the Disconfirmation condition exhibited greater parametric modulation in MPFC, LTJP, and the precuneus (see Table 1, Figure 3C).

These comparisons emphasize again the crucial importance of the social/informational context for consensus estimation. Even though participants showed *less* bias on average when presented with a social challenge in the Disconfirmation condition, it is also *only* in this condition that they exhibited coupling between activity in self, mentalizing, and valuation regions and actually observed bias.

DISCUSSION

In the present experiment, we sought to determine the neural correlates of trial-by-trial variation in observed consensus bias, as well as to assess whether or not these neural correlates would be sensitive to social context (specifically, whether or not there would be unique effects of social disconfirmation). The results show that activations in brain regions associated with self-related processes (MPFC BA10, precuneus), mentalizing (MPFC BA8/9, bilateral TPJ), and valuation (VMPFC BA11,VS) were strongly, positively associated with observed consensus bias, but *only* when participants were challenged by discrepant social feedback. In contrast, the same regions did not show significant associations with bias when social feedback was consistent with participants' own attitudes (in the Confirmation condition) or when information about others' attitudes was not present (in the No Information condition).

These results demonstrate that the social and motivational context of consensus estimation strongly affects the neural correlates of bias. The specificity of the association of activity in the analyzed regions with observed bias (i.e. that it is limited to the Disconfirmation condition only) is also suggestive. Presumably, it is in the Disconfirmation condition that participants have the strongest incentive to reassert the majority status of their attitudinal

positions, and the neural correlates of bias observed in this condition are consistent with motivated processing as a mechanism for sustaining consensus bias in the face of challenge.

In contrast, it is more difficult to explain the observed pattern of results if motivated processes are not involved. Of course, social projection might occur almost automatically if participants use their own attitude as a default or if selective exposure has led to biased sampling of attitudes from the broader population. However, if construed as impartial social thinkers (i.e. uninfluenced by motivations to defend their attitudes), participants should integrate discrepant attitudinal information into their consensus judgments. Mentalizing and self-related processes might play an invaluable role in this process, reconciling past knowledge with present feedback and yielding revised consensus estimates, and thereby *reducing* bias. If this were the case, we would expect activity in regions implicated in self-related cognition and mentalizing to be inversely associated with consensus bias following disconfirmation. Instead, we observe exactly the opposite. To the extent that an individual recruits these regions during consensus estimation, in response to disconfirmation, he or she is likely to show *enhanced* bias. Rather than impartially updating consensus estimates, it seems plausible that mentalizing and self-related processes are biased by the motivation to reassert the majority status of one's own attitudinal position.

What might be the mechanism by which mentalizing is biased? While the present results cannot adjudicate this issue, prior work on mentalizing suggests some possible avenues through which mentalizing mechanisms could be linked with motivations in ways that yield biased consensus estimates. For instance, mentalizing might be connected to biased retrieval and selection of attitude-relevant social knowledge. In work by Satpute, Badre, & Ochsner (2014), participants recruited mentalizing regions to a greater extent when the task demanded selection of goal-relevant social knowledge and the suppression of irrelevant information. In the present

context, mentalizing activity might therefore index a biased selection of social knowledge relevant to the goal of making self-serving consensus estimates (e.g. knowledge about the attitudes of affirming peers). Another intriguing possibility concerns the potential role of mentalizing (as well as self and valuation) regions in encoding social prediction error, both regarding others' outcomes (i.e. vicarious reward) and their expected actions. Joiner and colleagues (Joiner et al., 2017) review the extant literature on common and divergent neural correlates of prediction and error for self and for other, suggesting that encoding of social prediction error is not limited to the striatum and VMPFC, but includes a diversity of other areas as well, notably mentalizing regions. In the present study, the Disconfirmation condition violates the expectancy of participants, at least insofar as they expect others' attitudes to be similar to their own. It is therefore plausible that the recruitment of these regions is elicited by expectancy violation. However, it is our opinion that a motivational account of the subsequent processes subserved by these mentalizing regions best explains why activation in these regions is associated with enhancement, rather than attenuation, of consensus bias.

The results of this investigation are also generally consistent with an account of consensus bias that emphasizes the role of cognitive overlap between representation of the self and others. The positive association between trial-by-trial variation in MPFC activity and observed consensus bias might also therefore be interpreted in line with the literature in social neuroscience on the shared mechanisms involved in mentalizing and self-related cognition. The self may serve as an implicit anchor for mentalizing processes, from which we distance ourselves only effortfully. However, overlap between representations or self and other, by itself, does not explain the specificity of the effects to the Disconfirmation condition, in which participants presumably have the most reason to represent self and other distinctly. In one relevant study

(Tamir & Mitchell, 2010) elevated activity in DMPFC was associated with the dissimilarity between one's attitudes and those of another person, and VMPFC differentiated between items on which the self and the other person agreed and disagreed. The present research also reports elevated activity in DMPFC and VMPFC in response to social disconfirmation, but this increased hemodynamic response is parametrically associated with persistence in biased consensus estimates. Therefore, neither self-other overlap nor effortful distancing of the self from another seem (by themselves) to be a straight-forward explanation of the involvement of medial prefrontal regions in consensus bias. Instead, greater recruitment of DMPFC and VMPFC may be necessary precisely because, in the Disconfirmation condition, the attitude of a peer is presented to the participant as discrepant or distanced from one's own attitude. Participants may therefore need to do more motivated, effortful mentalizing work in order to bring representations of others' attitudes in line with their own. Hemodynamic association with consensus bias may therefore reflect this added effort involved in harmonizing own and others' opinions.

The results of the present study cohere most strongly with recent work on motivated cognition in the processing of social attitudes. For example, Hughes & Beer (2013) found that activity in the medial orbitofrontal cortex (MOFC), when under social-evaluative threat, was positively associated with participants' tendency to judge themselves to be "better-than-average" on a variety of personality traits. Thus, when challenged by the threat of social evaluation, participants' recruitment of brain regions involved in computations of subjective value resulted in (possibly biased) self-enhancement. Moreover, in a study by Kaplan, Gimbel, & Harris (2016) MPFC activity was associated with belief persistence in the face of challenge. In this study, participants who recruited MPFC most strongly when their attitudes were challenged showed the least change (or the most belief persistence), suggesting that the involvement of motivational

processes related to the self may have buffered their attitudes against social influence or conformity pressures. Importantly, however, their analysis does not examine consensus estimates regarding *others'* beliefs, nor does it examine within-subjects variation in MPFC activity. Whether or not one ought to change *one's own attitudes* in light of discrepant or contradictory feedback from others is a question fraught with moral and political significance; at least, it is not obvious that one ought always to align one's attitudes with the group consensus. But clearly we should at least be willing to consider discrepant feedback in forming our estimates of consensus. In this light, it is intriguing that neural mechanisms mediated by similar regions seem to insulate both our attitudes from social influence *and* our consensus estimates from up-date or revision.

Intriguingly, however, the results of the present work also resemble those of studies in which group consensus *affirmed* participants' antecedent attitudes, or in which participants actively brought their beliefs in line with group consensus. For example, Nook & Zaki (2015) found elevated reward-related activity in the nucleus accumbens when group consensus agreed with participants' preferences regarding food items, and VMPFC activity correlated positively with subsequent shifts in participants' ratings towards conformity with the group. Izuma and Adolphs (2013) observed that activity in the DMPFC mediated attitudinal shifts towards the opinions of positively-evaluated (liked) groups and away from negatively-evaluated (disliked) groups. These findings suggest that common processes may ultimately underlie consensus bias and social influence processes, despite the apparent divergence between paradigms and the opposed effects on attitude change. These processes may assist in the pursuit of consensus with the broader groups of which we are members – whether that consensus is veridical or not. When abundant and reliable information is available regarding group attitudes, we may be motivated to align ourselves with the priorities and values to which the group adheres. However, when

information about group consensus is more limited or ambivalent, the same desire to achieve congruence with the group may result in motivated misperceptions of group attitudes.

For example, in Nook & Zaki (2015), participants were lead to believe that social feedback represented an aggregate of the preferences of 200 fellow Stanford undergraduates. Resisting social influence in such a situation would therefore imply persisting in preferences that are strongly counter-normative, at odds with a large and (presumably) reliable sample of the attitudes of their peers. In the present study, UCLA participants believed that the social information merely reflected the attitude of *one* peer, making it much easier to discount this feedback if participants were motivated to maintain a contrary position. Interestingly, the quantity of social feedback available regarding others' attitudes and behavior was an important variable manipulated in Krueger and Clement (1994): participants in one experiment were presented with progressively larger samples of their peers who *unanimously* disagreed with their position. At all feedback sample sizes, participants' own position continued to significantly influence consensus estimates, but when the sample size of discrepant peers reached 20 participants no longer presumed that they were in the majority. Taken together, these results suggest an interesting hypothesis for future research: when participants believe that social information or social feedback accurately reflects the majority position of their peers, motivated processes operate to bring participants' positions in line with the group – whereas when participants believe that social information or social feedback may not reflect the majority position of their peers, similar motivational processes operate to bring consensus estimates in line with one's *own* position (and resist influence).

On our view, participants persist in exhibiting biased consensus estimates in the face of challenging feedback because the motivation to defend the social normativity of one's attitudes

biases mentalizing. Such an account nicely explains why, even though participants were *less* biased, on average, when forced to integrate in their consensus estimates attitudes opposed to their own, nevertheless, this same condition resulted in the *strongest* neural associations with bias. Importantly, strong consensus bias was also observed when participants' attitudes were confirmed and in the absence of social information. Other, presumably non-motivational, mechanisms must contribute to consensus bias in these conditions and future research will be needed to clarify their neural mechanisms. The present results should not be construed as excluding non-motivational causes of bias, but rather as evidence for motivated processing as an explanation for the persistence of bias in the face of disconfirmation.

We should note that the present paradigm does not rest principally on reverse inference, in which claims regarding psychological states are inferred from observations of brain activity. Instead, the conclusions drawn in this manuscript reflect a hypothesis testing approach. If motivated cognition influences consensus estimation, neural correlates of bias should differ in circumstances in which participants are motivated to defend the normativity of their attitudes (such as in instances of social disconfirmation). Results of both ROI and whole-brain analysis are consistent with this hypothesis, and inconsistent with the notion that identical mechanisms underlie consensus bias in both motivated and non-motivated contexts. Given the observed effects in regions implicated in self-related processes, mentalizing, and subjective valuation, we speculate that some confluence of relevant mechanisms biases the evaluation of others' attitudes and the integration of discrepant information with prior beliefs. However, we remain agnostic about the precise contribution of these regions as well as the exact means by which they combine to preserve bias in the face of disconfirmation. These topics must be addressed by future work, and the present data cannot adjudicate between competing mechanisms precisely.

Three crucial limitations of the present research might guide further efforts to examine the neural mechanisms supporting the false consensus effect. First, we did not scan participants while thinking about their own attitudes, so a direct comparison of neural responses when thinking about own and others' attitudes is not possible. Second, while connectivity analyses are complicated by the event-related nature of the present paradigm, they could illuminate more clearly the interactions between motivational and mentalizing processes that, we contend, shape consensus bias. Third, we must be careful concerning issues of causality when assessing neuroimaging evidence. It is possible, for example, that biased responses elicit subsequent activity in the observed regions, but that this activity does not cause bias directly. Future work might profitably clarify each of these issues and thereby offer a more complete picture of the consensus estimation process (and its tendencies towards bias) than is possible at present.

CONCLUSION

The present study examined the neural correlates of the false consensus effect in the presence and absence of social disconfirmation in the form of discrepant social attitudes from peers. The results revealed unique associations between trial-by-trial (issue-by-issue) consensus bias and hemodynamic activity in the Disconfirmation condition. In this condition, but not others, brain regions associated with self-related cognition (MPFC, precuneus), mentalizing (DMPFC, TPJ), and subjective valuation (VMPFC, VS) exhibited parametric modulation by the level of observed consensus bias, such that participants recruited these areas more strongly when they over-estimated support for their own attitudes (and under-estimated support for the opposing attitudinal position). The specific recruitment of these regions during social disconfirmation is consistent with an important role for motivated processing in sustaining consensus bias against

disconfirmatory feedback from others. These results shed light on the psychological and neural processes underpinning motivational sources of consensus bias in human social reasoning.

REFERENCES:

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

Beer, J.S., & Hughes, B.L. (2010). Neural systems of social comparison and the "above-average" effect. *NeuroImage, 49,* 2671-2679.

Brett, M., Anton, J-L., Valabregue, R., and Poline, J-B. (2002). Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in Neuroimage 16(2).

Clement, R.W., Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology*, *38*, 219-231.

Corbetta, M., Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3,* 201-215.

Crano, W. (1983). Assumed consensus of attitudes: the effect of vested interest. *Personality and Social Psychology Bulletin, 9,* 597-608.

Falk, E.B., Morelli, S.A., Welborn, B.L., Dambacher, K., & Lieberman, M.D. (2013). Creating buzz: the neural correlates of effective message propagation. *Psychological Science, 24,* 1234-1242.

Hughes, B.L., & Beer, J.S. (2012). Medial orbitofrontal cortex is associated with shifting decision thresholds in self-serving cognition. *NeuroImage, 61,* 889-898.

Hughes, B.L., & Beer, J.S. (2013). Protecting the Self: the effect of social-evaluative threat on neural representations of the self. *Journal of Cognitive Neuroscience, 25, 4,* 613-622.

Hughes, B.L., Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19(2)*, 62-64.

Izuma, K., Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron*, *78*, 563-573.

Joiner, J., Piva, M., Turrin, C., & Chang, S.C.W. (2017). Social learning through prediction error in the brain. *npj Science of Learning*, *2*, 1-8.

Kaplan, J.T., Gimbel, S.I., Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific Reports*, *6*, 1-11.

Krueger, J., Clement, R.W. (1994). The truly false consensus effect: an ineradicable egocentric bias in social perception. *Journal of Personality and Social Psychology, 67,* 596-610.

Krueger, J., Clement, R.W. (1997). Estimates of social consensus by majorities and minorities: the case for social projection. *Personality and Social Psychology Review*, *1(4)*, 299-313.

Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds). *Handbook of Social Psychology* (5th ed.) (pp. 143-193). New York, NY: McGraw-Hill.

Marks, G., & Miller, N. (1987). Ten years of research on the false consensus effect: an empirical and theoretical review. *Psychological Bulletin, 102,* 72-90.

Molenberghs, P., Johnson, H., Henry, J.D., Mattingley, J.B. (2016). Understanding the minds of others: a neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *65*, 276-291.

Morrison, K.R., & Matthes, J. (2011). Socially motivated projection: need to belong increases perceived opinion consensus on important issues. *European Journal of Social Psychology, 41,* 707-719.

Mullen, B., Hu, L.-T. (1988). Social projection as a function of cognitive mechanisms: Two meta-analytic integrations. *British Journal of Social Psychology*, *27*, 333-356.

Mullen, B., Smith, C. (1990). Social projection as a function of actual consensus. *The Journal of Social Psychology*, *130(4)*, 501-506.

Murray, R.J., Schaer, M., Debané, M. (2012). Degress of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self- and other-reflection. *Neuroscience and Biobehavioral Reviews*, *36*, 1043-1059.

Nook, E.C., Zaki, J. (2015). Social norms shift behavioral and neural responses to foods. *Journal of Cognitive Neuroscience*, *27(7)*, 1412-1426.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., Panksepp, J. (2006). Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *NeuroImage*, *31(1)*, 440-457.

Pronin, E., Gilovich, T., & Ross, L. (2004). Divergent perceptions of bias in self versus others. *Psychological Review, 111,* 781-799.

Pronin, E., Lin, D.Y., & Ross, L. (2002). The bias blind spot: perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28,* 369-381.

Raichle, M. (2015). The brain's default mode network. *Annual Reviews of Neuroscience*, *38*, 433-447.

Ross, L., Greene, D., & House, P. (1977). The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology, 13,* 279-301.

Satpute, A.B., Badre, D., & Ochsner, K.N. (2014). Distinct Regions of Prefrontal Cortex Are Associated With Controlled Retrieval and Selection of Social Information. *Cerebral Cortex*, *24*, 1269-1277.

Sherman, S.J., Presson, C.C., & Chassin, L. (1984). Mechanisms underlying the false consensus effect: the special role of threats to the self. *Personality and Social Psychology Bulletin, 10,* 127-138.

Sherman, S.J., Presson, C.C., Chassin, L., Corty, E., Olshavsky, R., (1983). The false consensus effect in estimates of smoking preference: underlying mechanisms. *Personality and Social Psychology Bulletin, 9,* 197-207.

Tamir, D.I., & Mitchell, J.P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of the Sciences of the United States of America, 107,* 10827-10832.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185,* 1124-1131.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, *30*, 829-858.

Wang, Y., Collins, J.A., Koski, J., Nugiel, T., Metoki, A., Olson, I.R. (2018). *Proceedings of the National Academy of the Sciences of the United States of America, 114,* E3305-E3314.

Welborn, B.L., Gunter, B.C., Vezich, I.S., Lieberman, M.D. (2017). Neural correlates of the false consensus effect: evidence for motivated projection and regulatory restraint. *Journal of Cognitive Neuroscience*, *29(4)*, 708-717.

Wolfson, S. (2000). Students' estimates of the prevalence of drug use: evidence for a false consensus effect. *Psychology of Addictive Behaviors, 14(3),* 295-298.

AUTHOR CONTRIBUTIONS:

B.L.W. and M.D.L. conceived and designed this research. B.L.W. conducted research and analysis. B.L.W. wrote the manuscript, with input and revisions from M.D.L.

COMPETING INTERESTS:

The authors declare no competing interests.

FUNDING:

Funding for this research was generously provided by Department of Defense 13RSA281 (PI: MDL).

ETHICS COMMITTEE:

Research described in the manuscript was approved by the North General Institutional Review Board of the University of California, Los Angeles.

Table 1:
Regions exhibiting parametric modulation by observed Consensus Bias

Test Effect/Anatomical Region	t	x	y	z	k
All Trials, Bias PM (+):					
Medial prefrontal cortex	6.06	-3	56	16	392

Ventromedial prefrontal cortex	4.74	9	62	-8	
	4.27	-9	59	-17	
<u>All Trials, Bias PM (-):</u>					
None					
<u>Confirmation, Bias PM (+/-):</u>					
None					
<u>Disconfirmation, Bias PM (+):</u>					
Left precuneus	7.90	-9	-52	4	588
	6.14	-18	-70	16	
	4.67	-9	-55	28	
Medial prefrontal cortex	5.22	-6	59	-8	967
	5.07	-6	62	22	
	4.90	9	50	7	
Left temoro-parietal junction	4.90	-42	-64	25	138
Thalamus	4.80	0	-10	7	294
Right amygdala	4.49	18	-7	-14	
Right caudate	4.37	18	23	4	
Left temporal pole	4.57	-66	-10	-17	178
	4.44	-54	11	-17	
Left inferior frontal gyrus	4.33	-36	32	-11	
<u>Disconfirmation, Bias PM (-):</u>					
None					
<u>No Information, Bias PM (+/-):</u>					
None					
<u>Non-Disconfirmation (Confirmation and No Information, Bias PM (+):</u>					
Left superior parietal lobule	5.94	-33	-49	70	82
	4.30	-21	-58	52	
<u>Non-Disconfirmation (Confirmation and No Information, Bias PM (-):</u>					
None					
<u>Disconfirmation Bias PM > Other Trials (Confirmation and No Information) Bias PM:</u>					
Left precuneus/PCC	5.57	-21	-70	13	776
	4.85	-6	-31	10	
	4.83	-9	-52	4	
	3.04	-3	-64	22	
Left temporo-parietal junction	4.83	-45	-67	28	216
	3.72	-36	-76	49	
	3.41	-63	-55	28	
Left caudate	4.59	21	-7	22	195
	4.26	3	-10	13	
	4.09	-3	5	-2	

Medial prefrontal cortex	4.48	-3	62	31	227
	3.83	39	62	10	
	3.68	3	65	10	

The above table displays whole-brain parametric modulation analyses, indicating regions in which hemodynamic activity covaried trial-by-trial with observed consensus bias. Results are reported both across all trials (ignoring information condition) as well as separately for trials within the Confirmation, Disconfirmation, and No Information conditions. Only grey-matter voxels were analyzed. Tabulated results are corrected for multiple comparisons, cluster FWE $p<0.05$, with a cluster-formation threshold of $p<0.005$. Peaks reported are separated by at least 20mm.

Figures:

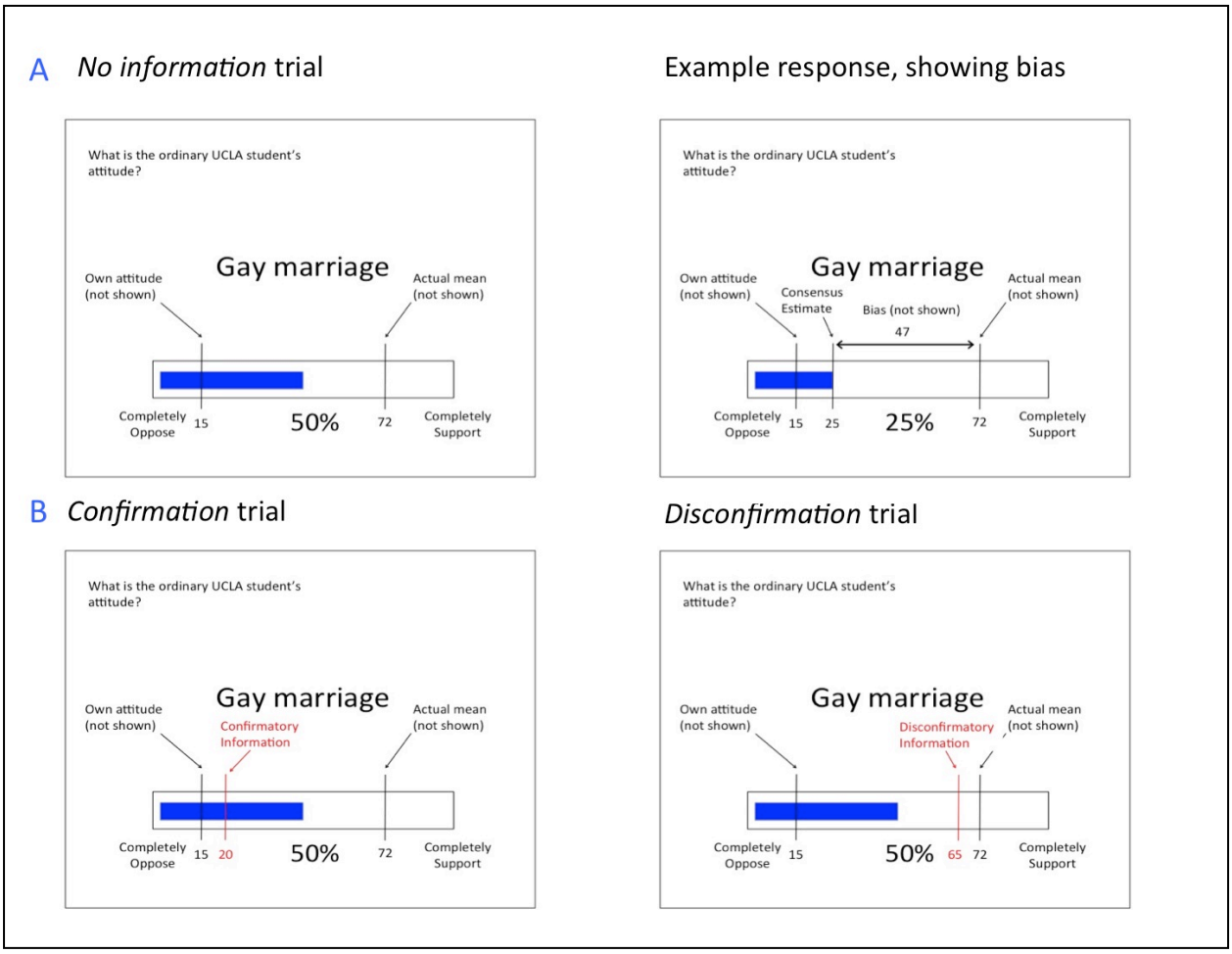

Figure 1: Depiction of trial structure and information presented on-screen. (A) The first panel shows an example screen for a No Information trial, in which a social or political attitude is presented to the participant for consensus estimation in the absence of any information ostensibly from the sample of UCLA undergraduates. In the second panel, a hypothetical response is depicted, in which a participant who opposes gay marriage selects a response that underestimates support for marriage equality in the undergraduate population. (B) Example trials from the Confirmation and Disconfirmation conditions. In the Confirmation condition, participants were presented with sample information suggesting that another undergraduate had an attitude similar to their own (no more than 5 points from their own attitude). In the Disconfirmation condition, participants were presented with sample information suggesting that another undergraduate had an attitude dissimilar to their own (at least 15 points different) and similar to the actual sample mean (within 5 points in either direction). These conditions were constrained by the experimental design to be exclusive, i.e., such that disconfirmatory information was always further from one's own attitude than confirmatory information, and always closer to the actual mean than the confirmatory information (see *Methods*).

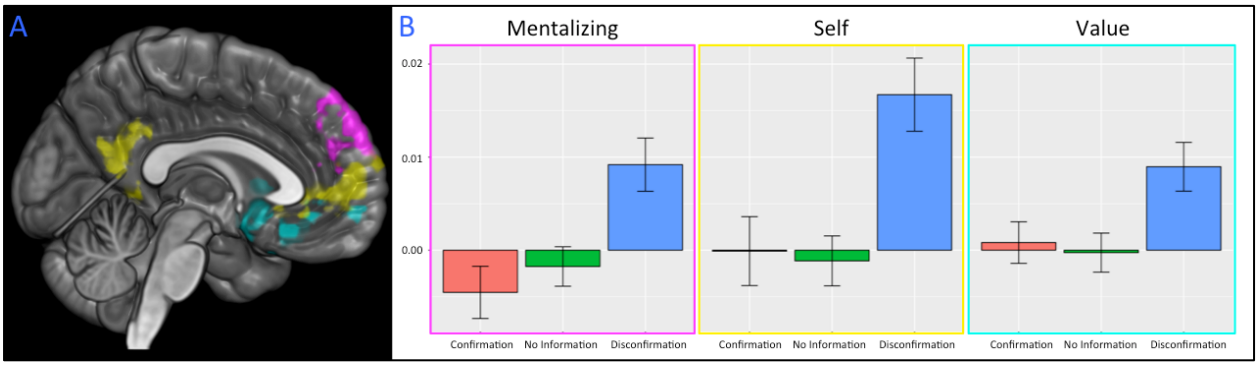

Figure 2: Regions of interest and associated parameter estimates from ROI analysis of parametric modulation of hemodynamic activity by observed consensus bias, across conditions. (A) Regions of interest (ROIs) for mentalizing (purple), self-related cognition (yellow), and subjective valuation (cyan), shown in mid-sagittal section. See Methods for detailed description of ROIs. (B) Parameter estimates for each of the ROIs plotted separately for each condition of interest. For all ROIs, the association between observed bias and hemodynamic activity is strongest in the Disconfirmation condition. Error bars indicate the standard error of the mean.

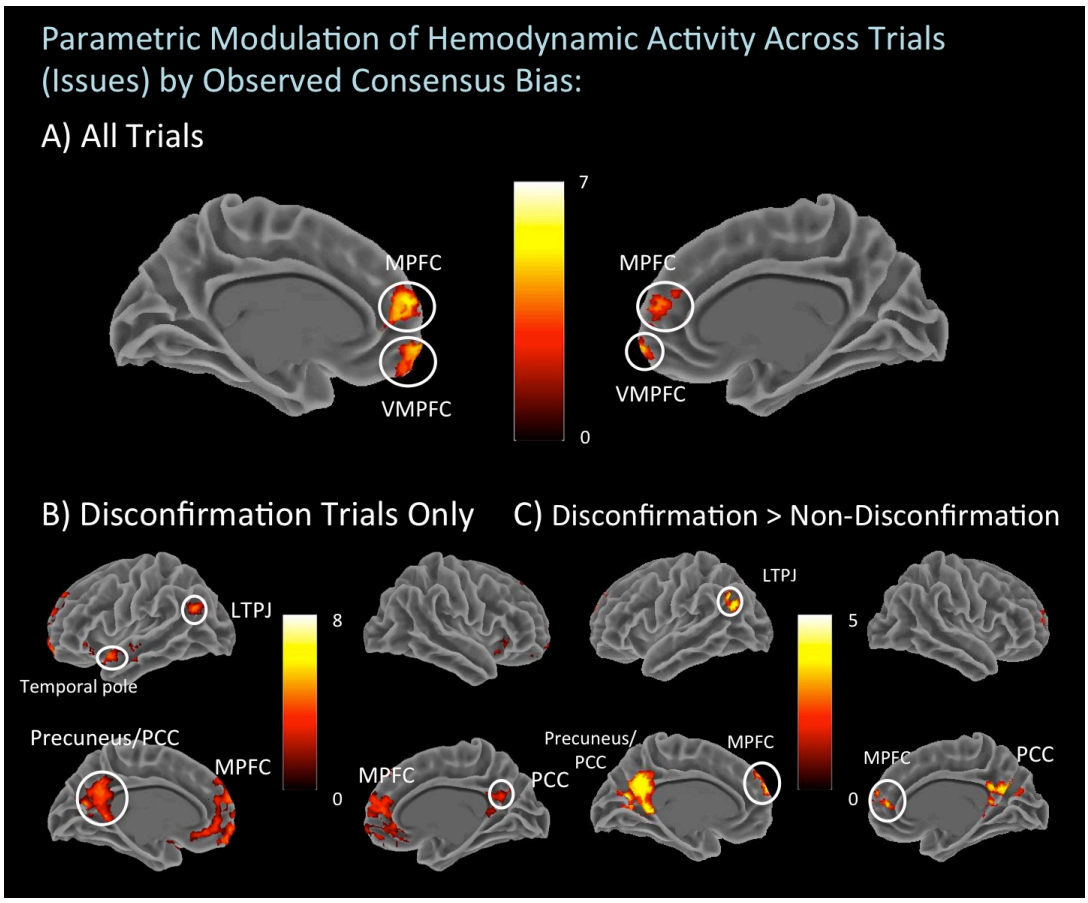

Figure 3: Parametric modulation of hemodynamic activity by observed consensus bias. Clusters are FWE corrected ($p<0.05$) with a cluster-formation threshold of $p<0.005$. See also Table 1. (A) Across all trials (ignoring condition), hemodynamic activity in medial prefrontal cortex (MPFC) and ventromedial prefrontal cortex (VMPFC) demonstrated parametric modulation by consensus bias, such that greater activity in these regions was associated with greater observed bias. B) For trials in the Disconfirmation condition (during which participants received social feedback discrepant with their own attitudes), positive associations with bias were observed both in regions implicated in mentalizing (DMPFC, left temporoparietal junction (LTPJ)), self-related cognition (MPFC, precuneus/posterior cingulate cortex(PCC)), and affective/motivational processes (amygdala (not shown), VMPFC). Stronger associations between consensus bias and activity were observed during the Disconfirmation trials than all other consensus estimation trials (i.e., Confirmation and No Information trials), as shown in (C).