Behavioral/Cognitive

# Brain Substrates of Recovery from Misleading Influence

**Micah G. Edelson,[1] Yadin Dudai,[1,4] Raymond J. Dolan,[2] and Tali Sharot[3]**

[1]Department of Neurobiology, Weizmann Institute of Science, Israel, Rehovot 7610001, [2]Wellcome Trust Center for Neuroimaging, Institute of Neurology, University College London, London, United Kingdom, WC1N 3BG, [3]Affective Brain Lab, Department of Cognitive Perceptual and Brain Science, University College London, London, United Kingdom, WC1E 6BT, and [4]Center for Neural Science, New York University, New York, New York 10003

**Humans are strongly influenced by their environment, a dependence that can lead to errors in judgment. Although a rich literature describes how people are influenced by others, little is known regarding the factors that predict subsequent rectification of misleading influence. Using a mediation model in combination with brain imaging, we propose a model for the correction of misinformation. Specifically, our data suggest that amygdala modulation of hippocampal mnemonic representations, during the time of misleading social influence, is associated with reduced subsequent anterior–lateral prefrontal cortex activity that reflects correction. These findings illuminate the process by which erroneous beliefs are, or fail to be, rectified and highlight how past influence constrains subsequent correction.**

*Key words:* **brain; fMRI; memory; recovery; social**

## Introduction

From early infancy, we look to others as a primary source of information about the world. This reliance is so powerful that we often reevaluate our own perceptions, preferences, and memories when they contradict a larger consensus (Sherif, 1936; Meade and Roediger, 2002; Hirst and Echterhoff, 2012; Lewandowsky et al., 2012). Although this strategy can often be adaptive in maximizing accuracy, because other members of the group may have more accurate knowledge than the individual (Deutsch and Gerard, 1955; Surowiecki, 2004; Schacter et al., 2011), such dependence may carry a cost when relying on noncredible sources: individuals with inaccurate information, poor skills, or people who are intentionally lying. In such situations, it is adaptive to surmount the misleading influence and maintain fidelity to the original mnemonic representation (Byrne and Whiten, 1989; Schiller et al., 2008; Schacter et al., 2011; Lewandowsky et al., 2012; Engelmann and Hein, 2013).

This process, however, is not always successful (Ross et al., 1975; Braun and Loftus, 1998; Meade and Roediger, 2002; Echterhoff et al., 2005; Skurnik et al., 2005; Lewandowsky et al., 2012). For example, eyewitnesses can often be influenced by other witnesses, leading to a testimony that differs from their original experience (Wright et al., 2009; Hirst and Echterhoff, 2012; Schacter and Loftus, 2013). If they subsequently discover that the source they relied on had low credibility, they are not always able to recover from such influence and reclaim their original beliefs (Chambers and Zaragoza, 2001; Meade and Roediger, 2002; Echterhoff et al., 2005). Additionally, misinformation conveyed by medical professionals has been demonstrated to have long-lasting effects on individuals, even after they are informed it was mistaken (Lewandowsky et al., 2012). Certain forms of advertising (Braun and Loftus, 1998; Skurnik et al., 2005; Lewandowsky et al., 2012) and political propaganda have similar effects (e.g., as in the case of the controversy over Barack Obama's birthplace; Lewandowsky et al., 2012). How restoration from misleading influence takes place in the brain and what are the brain processes that restrict such recovery, even when the original source of influence is discredited, remain unanswered.

We posited that the ability to correct past influence depends on brain processes occurring at two temporally distinct phases: (1) the time of exposure to influence (initial influence strength) and (2) the time influence is lifted. We have previously demonstrated that amygdala activation during exposure to social information and its enhanced functional connectivity with the hippocampal-dependent memory system reflect robust social influence (Edelson et al., 2011). Recovery from misleading social influence can be taken as a powerful example of revision of former beliefs. Here, we first identified brain activity that is related to correction when the source of the influence is discredited. We then conducted a mediation analysis (Hayes, 2013) to test whether and how this activity and correction success were modulated by past activity related to social influence.

Received Nov. 7, 2013; revised April 15, 2014; accepted April 21, 2014.

Author contributions: M.G.E. and T.S. designed research; M.G.E. performed research; M.G.E. and T.S. analyzed data; M.G.E., Y.D., R.J.D., and T.S. wrote the paper.

M.G.E. and Y.D. were supported by a Weizmann Institute–United Kingdom Making Connections Grant. T.S. was supported by a Wellcome Trust Career Development Fellowship. R.J.D. was supported by Wellcome Trust Senior Investigator Award 098362/Z/12/Z and a Wellcome Trust Strategic Award 091593/Z/10/Z. Y.D. was supported by the Center of Research Excellence in the Cognitive Sciences of the Planning and Grants Committee and Israeli Science Foundation (Grant 51/11) and by the EP7 Human Brain Project. We thank E. Phelps, J.G. Edelson, A. Ben-Yakov, L. Pell, K. Ludmer, T. Fitzgerald, S. Fleming, A. Mendelsohn, A. Pine, D. Kumaran, and N. Wright for helpful comments and the support teams of the Norman and Helen Asher Center for Brain Imaging at the Weizmann Institute and the Imaging Neuroscience & Theoretical Neurobiology group at the Wellcome Trust Center for Neuroimaging at University College London.

The authors declare no competing financial interests.

This article is freely available online through the *J Neurosci* Author Open Choice option.

Correspondence should be addressed to Dr. Micah Edelson, Department of Neurobiology, Weizmann Institute of Science, Rehovot 7610001, Israel. E-mail: micah.edelson@weizmann.ac.il.

Copyright © 2014 Edelson et al.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

## Materials and Methods

*Participants.* Fifty-nine right-handed participants (28 females, average age 28.1 ± 1.0 years) participated in the study. One participant was excluded because of suspected minor brain pathology, one was excluded because of head movements exceeding 4 mm, and one participant's data could not be acquired because of claustrophobia in the scanner setting. Three additional participants were excluded from analysis because their imaging data were not acquired (because of technical malfunction in the scanner). Only subjects who indicated no suspicion of the experimental manipulation when debriefed were included in the analysis (final $N = 40$; 19 females, average age 27.8 ± 1.1; 20 in the Social group and 20 in the Computer group; groups are defined below). The percentage of suspecting subjects was well within the norm for conformity studies (for a summary of suspicion in conformity studies see Stang, 1976; Ortmann and Hertwig, 2002). All participants gave informed consent and were remunerated for participation. The study was approved by the Institutional Review Board of the Sourasky Medical Center, Tel-Aviv.

*Stimuli.* The stimuli consisted of a 40 min eyewitness-styled documentary following the activities of the police deporting undocumented immigrants. The film included scenes of forceful arrests of immigrants and their families. The content was tested previously and had medium emotional valence as rated by participants (51.2 ± 2.3 on a 0–100 scale; Edelson et al., 2011).

*Procedure.* The experiment was divided into four phases (encoding followed by memory Tests 1–3) (see Fig. 1). Two groups (Social and Computer) underwent a similar protocol that differed in the source of the manipulation (in Test 2; see below).

*Encoding phase (day 0).* The initial encoding of the movie was performed in groups of five unacquainted individuals (see Fig. 1*A*). The participants were introduced to the group, and a photograph was taken of each participant. The subjects were told that the experiment was testing memory and they would subsequently be asked questions concerning the content of the film. They were specifically instructed not to discuss the film or memory tests with others at any stage.

*Memory Test 1: initial memory, outside the scanner (day 3).* Memory Test 1 served to assess the participants' baseline memory accuracy and confidence before the manipulation and was comprised of a computerized 400 question multiple-choice memory questionnaire on the film's content (see Fig. 1*B*). After selecting one of the two possible answers on each question, the subjects rated how confident they were in their responses on a visual analog scale (VAS) ranging from 0 (guess) to 100 (absolute confidence) with 25 equating to low confidence, 50 to medium confidence, and 75 to high confidence. For the Social group, the average accuracy in this test was 69.1 ± 1.2% for all answers and 80.2 ± 2.0% for answers with medium to high confidence scores. For the Computer group, the average accuracy was 67.0 ± 1.9% for all answers and 82.3 ± 4.1% for answers with medium to high confidence scores.

*Memory Test 2: Manipulation phase, in the scanner (day 7).* Here an attempt was made to influence participants' answers (see Fig. 1*C*).

*Social manipulation group (Social group).* Thirty participants (12 females, age 28.6 ± 0.8) answered the same memory questions as in Test 1, but this time before answering, they were presented with answers supposedly given by their four fellow co-observers. On each trial, the pictures of the co-observers were displayed together with their supposed answers (mode of presentation adapted from Berns et al., 2005). The participants then selected their answers and on 66% of the trials also provided a confidence rating. Participants were instructed that the answers of their co-observers could be used to assist their retrieval but that they ultimately were required to answer according to their own recollection. To minimize scanning time, only 320 questions were included in this test [i.e., 80 Credibility questions (see definition below) were randomly excluded].

The co-observer answers were pseudo-randomly allocated into three different categories as follows (see Fig. 1*C*).

*1. Manipulation condition.* In a subset of trials, for which the participants originally had a confident veridical memory (as identified by Test 1), the answers provided by the four co-observers were all incorrect. For each subject, questions that were answered correctly by that subject in memory Test 1, with a confidence rating from 70% to 140% of his/her average confidence rating, were identified. Eighty of these questions (randomly assigned) were included as Manipulation questions in Test 2. Average confidence rating for all Manipulation questions was 62.6 ± 2.3, lying between a medium (50) and high (75) confidence rating.

*2. No-Information condition.* Twenty-five questions were randomly chosen from the same pool of questions as in category 1 above (average confidence rating 62.5 ± 3.3). For these questions, the co-observers' answers were not made available, and instead the letter X was displayed.

*3. Credibility condition.* A total of 215 different questions were randomly chosen from all questions in memory Test 1. Because it is not credible that all co-observers' answers would be in disagreement with the participant's remembered view, it was necessary to add additional questions in which the co-observers' answers supported the participant's remembered view and confidence. The pattern of the falsified co-observer answers in this condition depended on the subject's answer and confidence in memory Test 1, such that the greater the subject's confidence in his/her answer, the greater the number of conforming answers were given by co-observers (i.e., between two and four co-observers' answers were in agreement with the subject's answer in Test 1). Specifically, for questions in which participants answered with low confidence in Test 1 (<25 in VAS, see Fig. 1, confidence scale), two co-observers were in agreement and two in disagreement with the participant's original answer, emphasizing the difficulty and uncertainty related to the specific question. For answers with medium confidence ratings [between low (25) and high (75)], three co-observers were in agreement and one in disagreement with the participant's original view, again strengthening the participant's original perspective. For answers with very high confidence (>75), all co-observers were in agreement with the participant's original answer. To further increase the protocol's credibility and emulate natural human variability, 15% of Credibility questions received a random answer pattern (i.e., regardless of the answer and confidence in Test 1, between two and four co-observers' answers were correct). The Credibility questions always provided partial or full support for the participant's original answer, whereas in the Manipulation questions co-observers' answers were always in unanimous disagreement with the participant's original answer. Previous research into the phenomenon of conformity and social influence demonstrated that the addition of even one confederate supporting the subject's view was sufficient to significantly reduce social influence, and when half of the group supported the subject's answer, the affect was completely abolished (Allen, 1975; Asch, 1951). The average confidence for Credibility questions was 63.9 ± 3.6, which was not different from the confidence for Manipulation questions (62.6; $t_{(38)} = 0.1$, $p = 0.9$).

*Computer manipulation group (computer group).* To compare recovery from social and nonsocial influence, we performed a control fMRI experiment using an inanimate medium to convey misinformation. In this control, 29 participants (16 females, age 27.6 ± 1.2 years; final $N = 20$ after exclusion) underwent a protocol similar to that of the main experiment. However, in the Manipulation phase, instead of receiving answers from co-observers, participants were told that the information originated from four different computer algorithms, a common technique used to control for social effects (Berns et al., 2005; Klucharev et al., 2009). Each computer algorithm was said to use a separate database of documentary films and to provide the most probable answer according to this database. The participants were told that the different algorithms have been tested and provide an accuracy level similar to that of humans. The average confidence rating was 60.0 ± 2.6 and 62.5 ± 2.5 for the Manipulation and No-Information conditions, respectively.

*Memory Test 3: correction phase; in the scanner (day 14).* To examine the neural and computational processes underlying recovery from social influence, participants were informed that the answers given by the co-observers/computers during the previous session were actually determined randomly (see Fig. 1*D*). This rendered these answers as "uninformative." The participants were then requested to complete the memory test again (Test 3) based on their original memory of the movie. The scan was divided into three sessions with a 15 min break between sessions. On each trial, the question was presented along with two possible answers for 1.5 s. The participants were not allowed to answer during this period to ensure that they read the question and answers fully. After

the 1.5 s interval, the color of the question font changed, indicating to the subjects that they could now respond. Participants then provided a response. This was followed by a fixation cross for a jittered time interval (3 s average). On 66% of the trials, the participants also provided their confidence rating. In these cases, an additional jittered fixation was introduced between the participants' response and confidence rating (2 s average).

*Debriefing.* After conclusion of the four phases of the experiment, participants were interviewed using a questionnaire in which they were asked whether in any stage of the Manipulation phase they suspected that the answers provided to them did not originate from the co-observers/computers. In accordance with previous work in experimental psychology, data from participants who suspected the nature of the manipulation were analyzed separately (Stang, 1976; Ortmann and Hertwig, 2002). Two questions were defined as predetermined criteria to determine suspicion. The questions were as follows: (1) Did you suspect a manipulation at any stage of Test 2? If so, at what stage? Excluding answer: Yes (regardless of the stage). (2) Did you think the answers of your fellow group members were fictitious? If so, how strong was your suspicion? Excluding answer: Yes (regardless of the strength of suspicion). All subjects that gave a positive answer indicating suspicion on one of the questions also answered positively, indicating suspicion on the other question. These subjects reported a gradual acquisition of suspicion over the test period. Thus, suspicious subjects were a heterogeneous group comprised of subjects that suspected the manipulation at various stages of the test and with varying degrees of suspicion. To avoid Type 1 errors and confounds related to uncovering the manipulation at various stages during the imaging session, we separately analyzed all suspecting subjects. Eight subjects in the Social group and five subjects in the Computer group were suspicious. A wealth of psychological literature demonstrates this number is well within the norm for conformity studies (for summary of suspicion in conformity studies see Stang, 1976; Ortmann and Hertwig, 2002). Behavioral and imaging analysis confirmed that suspecting subjects indeed differed from nonsuspecting subjects. Whereas memory performance in Test 1, before manipulation was introduced, did not differ between suspecting and nonsuspecting subjects for either the Social group or Computer group ($p > 0.2$), in Test 2 (when manipulation was introduced) suspecting subjects had significantly fewer errors on Manipulation trials (Social group: 48.6 ± 4.5%; Computer group: 26.6 ± 5.7%) than the nonsuspecting subjects (Social group: $t_{(26)} = 3.2$, $p < 0.005$; Computer group: $t_{(23)} = 2.2$ $p < 0.05$). The interaction between group (Social/Computer) and exclusion (excluded/included) for conformity levels was not significant, indicating that the decrease in conformity was evident for suspecting subjects in both groups ($F_{(1,49)} = 0.3$ $p = 0.6$). Moreover, no significant activations were found for the region of interest (ROI) identification contrast for excluded subjects (Manipulation trials > Credibility trials; Social and Computer groups), even when using a relatively low threshold (0.001 uncorrected; $k$ (cluster size) > 10). No significant correlation was found with the Change of Mind (COM) parameter in ROIs identified in main text. Because $N$ here is small, these null findings cannot be reliably interpreted.

*Image acquisition and statistical analysis.* All statistical tests reported are two-sided. All *post hoc* tests are Bonferroni-corrected for multiple comparisons, including correction for number of ROIs and reported $p$ values are multiplied to include this correction when appropriate. Variance is reported in SEM. Statistical analysis was performed using R (R Project for Statistical Computing, RRID:nif-0000-10474), MATLAB (2012 MathWorks; RRID:nlx_153890), and SPSS (version 21, IBM).

*Image acquisition.* Imaging was performed with a 3T Siemens Trio Magnetom scanner at the Ascher Imaging Center in the Weizmann Institute. All images were acquired using a 12-channel head matrix coil. Three-dimensional T1-weighted anatomical scans were acquired with high resolution 1 mm slice thickness (3D MP-RAGE sequence, TR 2300 ms, TE 2.98 ms, 1 $mm^3$ voxels). For BOLD scanning T2*-weighted images were acquired using the following parameters: TR 2000 ms, TE 30 ms, flip angle 80°, 35 oblique slices without gap, 30° toward coronal plane from AC PC, 3 × 3 × 4 mm voxel size covering the whole cerebrum.

*Image analysis.* Statistical Parametric Mapping (SPM8; Wellcome Trust Centre for Neuroimaging, London; http://www.fil.ion.ucl.ac.uk/spm) was used to analyze the fMRI data. After discarding the first three dummy volumes, images were realigned to the first volume, unwarped, normalized to a standard EPI template based on the MNI reference brain, resampled to 2 mm × 2 mm × 2 mm voxels, and spatially smoothed with an isotropic 8 mm full width at half-maximum Gaussian kernel.

*Reaction times (RTs).* An ANOVA for RT with condition type (Manipulation/Credibility) and group (Social/Computer) as factors revealed no significant interaction. These results suggest that RT differences cannot be a sufficient explanation for our results. Notwithstanding, the duration of each event was specified in the first-level analysis (Fleming et al., 2012).

*ROI identification.* All parameter estimates extracted from ROIs represent the average value across all voxels in the ROI.

*1. Functional ROIs.* To identify candidate regions participating in recovery from past influence, we constructed the general linear model (GLM) detailed below and performed an unbiased whole-brain contrast searching for regions where BOLD response was greater for all misinformation trials relative to all Credibility trials (Manipulation > Credibility) regardless of group membership (i.e., over both the Social and Computer groups together; whole-brain Family Wise Error (FWE) corrected; $p < 0.05$; cluster size ($k$) > 50). This resulted in five regions: anterior-lateral prefrontal cortex (alPFC), bilateral inferior parietal cortex (IPC), superior medial prefrontal cortex (smPFC), and superior lateral prefrontal cortex (slPFC). These regions were used as ROIs to constrain subsequent analyses that compared the extent of correction in the Social group versus the Computer group. The ROI identification does not bias later comparison of Social and Computer groups, as initial identification is conducted over both groups in an unbiased manner (Kriegeskorte et al., 2009). For completeness, we additionally report a parietal region that was active in both groups.

*GLM model for ROI identification (Test 3).* For each participant, a time series was created indicating the temporal position of the different trial types convolved with the canonical hemodynamic response using a random-effects GLM. The critical time window from question presentation to subject's response was modeled into the following: (1) No-Information condition trials, (2) Credibility condition trials, and (3) Manipulation condition trials. An additional regressor was created for the time window of the confidence rating phase. For the contrast approach model only (see Results), the Manipulation condition was further divided according to whether the subject gave correct/incorrect answers in Test 2 and Test 3.

Given that Credibility questions are comprised of questions initially answered correctly as well as questions initially answered incorrectly (Test 1) whereas the Manipulation questions are comprised only of questions initially answered correctly, we additionally performed the same ROI selection analysis as above but included only Credibility questions answered correctly in Test 1. This resulted in identification of the same ROIs (excluding the left IPS, FWE < 0.05). As an additional control, we added confidence scores on Test 3 as a covariate modulating the time between question presentation and subject's response. This resulted in identification of the same five ROIs as above (whole-brain corrected, FWE < 0.05 $k > 50$).

*2. Anatomically defined ROIs.* The *a priori* anatomical ROIs (left amygdala and left anterior hippocampus) were selected based on their involvement at the time of social influence (Edelson et al., 2011) and defined based on known anatomical landmarks according to the Talairach Daemon Atlas (Lancaster et al., 1997) using the SPM WFU PickAtlas tool (Maldjian et al., 2003).

*COM parametric analysis.* The COM value representing the subject's COM during the Correction phase was calculated for each Manipulation trial in which confidence rating was available on both Test 2 and Test 3 (average number of events per subject = 35 ± 0.7) using the equation below (see Results section for more information) as follows:

$$COM = (\alpha_{Test3} * Confidence\ Test\ 3) - (\alpha_{Test2} * Confidence\ Test\ 2)$$

$$For\ correct\ answer\ \alpha = 1;\ for\ incorrect\ answer\ \alpha = -1$$

COM values were calculated for each participant per each event. For the behavioral analysis (see Fig. 3), each participant's COM values were av-

eraged per condition (i.e., Manipulation, Credibility, No-Information) and then compared across conditions and groups (i.e., Social vs Computer). For the brain-imaging data, we created a regressor for each participant with one COM value per each Manipulation trial, modulating the BOLD signal from question presentation until response. The resulting parametric COM values were averaged per participant across voxels in each ROI and then compared across groups (Social vs Computer). This GLM was identical to the one described above for ROI identification except that, for increased statistical power, the three scanning sessions in Test 3 were concatenated and three constant terms were included to represent each session (see SPM manual; http://www.fil.ion.ucl.ac.uk/spm/doc/spm8_manual.pdf). Parametric regressors were automatically orthogonalized to main effect regressors (SPM8; Wellcome Trust Centre for Neuroimaging, London; http://www.fil.ion.ucl.ac.uk/spm). Absolute confidence levels during Test 3 were controlled for by adding these values as an additional regressor preceding the COM regressor in the first-level SPM model. Adding or removing this covariate did not significantly alter the results. More complex models of COM with a larger number of free parameters (up to 5 degree polynomial model) did not significantly improve our model fit over a simple linear model.

*Mediation analysis.* We created a mediation model for each subject linking past amygdala activation with subsequent COM (via activity in mediator regions; hippocampus and alPFC) per event. Thus, a GLM had to be constructed for which activation for each event during Test 2 and Test 3 could be extracted in the three ROIs and those values then entered into each subject's mediation analysis. Mediation parameters were then averaged across subjects. To that end, for each participant, a time series was created indicating the temporal position of each trial (one GLM was constructed for Test 2 and one for Test 3). Data for individual trials were convolved with the canonical hemodynamic response using a random effects GLM. For this GLM, each event was treated as a regressor, a technique used in a similar context (Charpentier et al., 2014) as well as in multivariate, functional connectivity, and mediation studies (Rissman et al., 2004; Atlas et al., 2010; Bonnici et al., 2012; Chadwick et al., 2012; Mumford et al., 2012; Cisler et al., 2014). For Test 2, parameter estimates were extracted from the left amygdala (lAmy) and left anterior hippocampus (laH) anatomical ROIs for each event separately during the Manipulation phase. The same analysis was done in the left alPFC ROI for Test 3. These values, for each event, were fed into the mediation model below. In comparison with previous imaging studies, creating parameter estimates per event allowed us to improve the statistical power of the mediation model that is limited by the typical amount of participants in imaging studies.

*Mediation model.* A mediation analysis was performed following the modern mediation format (Wager et al., 2008; Atlas et al., 2010; Hayes, 2013). The regions participating in the model were the anatomically defined left amygdala and left anterior hippocampus and the functionally defined left alPFC (see Fig. 4*A*). Using multiple regression models, we calculated within each subject the following regression parameter estimates (via R programing software and PROSESS SPSS macro, Model 6; Hayes, 2013; parameter estimate of interest marked in bold; $\varepsilon$ = error term).

1. A relationship between amygdala (initial predictor) and anterior hippocampus (first mediator) activations as follows: laH activation = **beta1.1** * lAmy activation + $\varepsilon$.
2. A relationship between laH (second mediator) and LalPFC (second mediator) when lAmy activation is included in the model as follows: LalPFC activation = **beta2.1** * laH activation + beta2.2 * lAmy activation + $\varepsilon$.
3. A relationship between LalPFC (first mediator) and COM (behavioral outcome) when both lAmy and laH activation are included in the model as follows: COM = **beta3.1** * LalPFC activation + beta3.2 * laH activation + beta3.3 * lAmy activation + $\varepsilon$.
4. Indirect effect of interest (amygdala influence on COM via the anterior hippocampus and alPFC mediators). For each subject, the indirect effect of interest is defined as the multiplication product of the aforementioned parameter estimates of interest (i.e., points 1 until 3) (Wager et al., 2008; Hayes, 2013).
5. Total and direct effects. As defined by Hayes, 2013, the total effect is the relationship between the amygdala (initial predictor) and COM (outcome) before controlling for the mediators (COM = **beta4.1** * lAmy$\alpha$ + $\varepsilon$). The direct effect is defined as the relationship between amygdala (initial predictor) and COM (outcome) after discounting the variance explained by the mediation (Hayes, 2013).

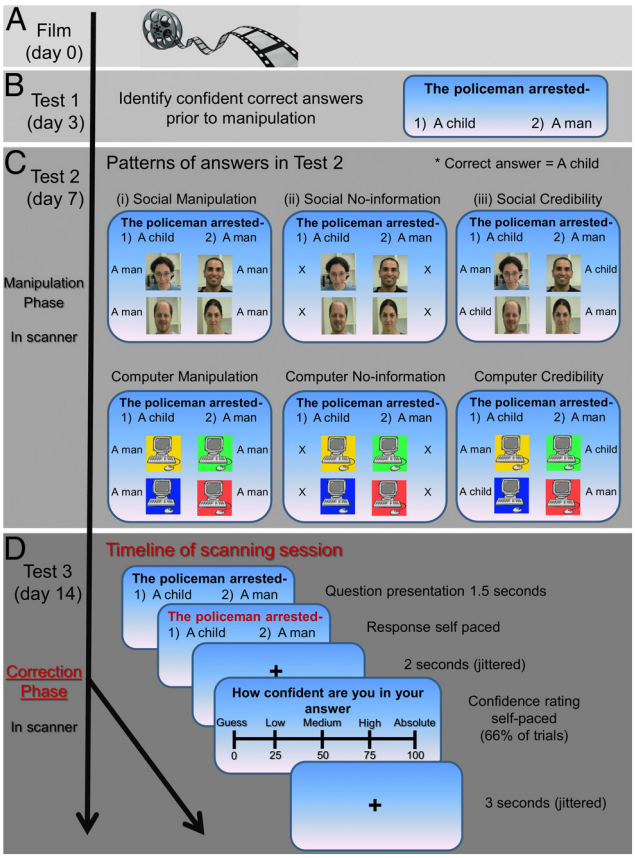


**Figure 1.** Experimental outline. ***A***, Subjects viewed a film in groups of five and subsequently performed three memory tests individually. ***B***, Test 1 assessed the participants' initial memory and confidence. ***C***, In the Manipulation phase (Test 2), either Social or inanimate (Computer) influence was induced. ***D***, In the Correction phase (Test 3), the influence was removed and recovery was examined. The Manipulation phase (***C***) had three different experimental conditions: (1) the Manipulation condition, in which all co-observers' answers were incorrect; (2) the No-Information condition, in which the letter X was displayed instead of co-observers' answers; and (3) the Credibility condition, in which variable patterns of co-observers' answers were displayed. Imaging data reported correspond to the Correction phase scanning session (***D***), in which each question and possible answers were presented for 1.5 s. Subsequently, a font color change indicated that the participants could respond. Finally, confidence ratings were provided on 66% of the trials.

All parameter values were then taken to a group level analysis. To make minimum assumptions on the distribution of regression products, group level analyses were conducted using a nonparametric Wilcoxon signed-rank test, and the resulting $z$ and $p$ values were reported (Howell, 1997). The use of a parametric test resulted in the same findings. To maintain reference to older mediation models, we additionally performed a Sobel test to determine whether the addition of the indirect pathway significantly reduced the direct pathway.

## Results

To examine whether and how people correct their beliefs following misleading social influence, it was critical to first demonstrate that participants were indeed influenced by the false information. Indeed, when presented with unanimous incorrect judgments of co-observers (Social Manipulation trials, Test 2; Fig. 1*Ci*), partic-

ipants followed the false opinion of the majority in 68.3% of the trials (as we have previously reported in Edelson et al., 2011). This was despite providing accurate confident responses to the same questions during Test 1 (see Materials and Methods). Such conversion to erroneous judgments was not explained by simple forgetting because the conversion rates in this condition were significantly greater than when participants were not presented with answers of co-observers at all (Fig. 1*Cii*; mean conversion rate of No-Information condition = 15.5%; significantly lower than Social Manipulation condition; $t_{(19)} = 16.9$; $p < 10^{-5}$; here and subsequently, $p$ values are Bonferroni-corrected), or when presented with accurate/mixed responses of co-observers (Fig. 1*Ciii*; mean conversion rate of Credibility condition = 11.8%; significantly lower than Social Manipulation condition, $t_{(19)} = 19.7$; $p < 10^{-5}$). Furthermore, error rates were greater when unanimous false information was delivered by humans rather than computers (Computer group Manipulation condition = 52.7%; significantly lower than Social Manipulation condition, $t_{(38)} = 2.7$; $p < 0.05$). Entering conversion rates into a 2 × 3 ANOVA with group (Social/Computer) and condition (Manipulation/No-Information/Credibility) as factors revealed a significant interaction between group and condition, as well as a main effect for condition and group (respectively: $F_{(2,76)} = 5.8$; $F_{(2,76)} = 207.3$; $F_{(1,38)} = 4.9$; $p < 0.05$). These results indicate that participants altered their responses to conform to the erroneous judgments of others (Edelson et al., 2011). The current work focuses on how participants corrected these errors once they learned they were misguided.

**Correcting for past social influence: behavioral manifestation**

When participants learned they were misguided, they corrected past errors on 61% of the error trials that were due to Social Manipulation (equal to 40.4% of all Social Manipulation trials) and 59% of error trials due to the Computer Manipulation (31.6% of all trials in this condition). This was not due to random fluctuations in judgment because these numbers were significantly lower on Credibility and No-Information trials ($t_{(19)} = 12.6$ and $t_{(19)} = 14.5$ respectively; $p < 10^{-5}$).

To quantify the extent to which a belief is altered, it is important to take into account not only the change in judgment but also changes to the confidence in that judgment. Successful recovery is possible because we maintain a representation of the original correct information, even after the creation of a new competing erroneous representation (Lewandowsky et al., 2012). Such competing representations can be continually compared and interchanged depending on the evidence supporting each representation (Vickers, 1970; Bogacz et al., 2006; Vlaev et al., 2011; De Martino et al., 2013), which in turn can lead to change of mind (Resulaj et al., 2009). Discrediting social influence, in this context, introduces new evidence that can shift the balance toward the original correct representation. Prominent neural computational models (such as drift diffusion models) postulate that, in making a selection between two competing options, the brain accumulates noisy evidence supporting each option until one option reaches a threshold (Vickers, 1970; Bogacz et al., 2006; Resulaj et al., 2009; Vlaev et al., 2011; De Martino et al., 2013). Confidence in the selection then represents, at the point of choice, the strength of support for one option compared with the alternative (De Martino et al., 2013). As illustrated in Figure 2, shifting from a low confidence erroneous judgment (Xi) to a low confidence correct judgment (Xj; case I) represents less of a change than shifting to a high confidence correct judgment (Xk; case II). Furthermore, a subject may maintain the same answer but alter his/her confidence in that answer. For example, one may

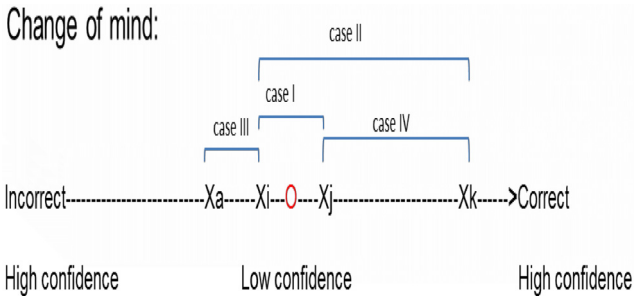


**Figure 2.** COM flowchart model illustration, spanning from an incorrect answer with high confidence to a correct answer with high confidence. According to the model, a larger shift on this axis after influence is removed corresponds to a larger COM.

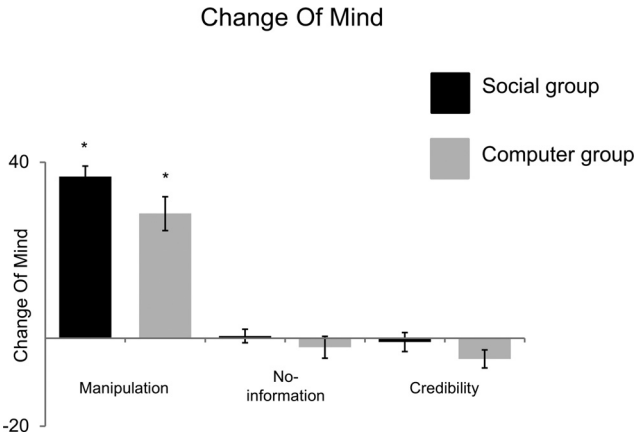


**Figure 3.** Behavioral manifestation of reversal. COM experimental results. The parameter representing the change in confidence toward the correct answer between the Manipulation phase (Test 2) and Correction phase (Test 3), for Manipulation, No-Information, and Credibility events. *$p < 10^{-5}$.

have lower confidence in an incorrect answer after misinformation is discredited (i.e., case III; move from Xa to Xi), revealing a change in the strength of the judgment. Or a subject may stick with a correct answer in the face of social influence but have enhanced confidence in this correct answer after learning that the other participants' false answers were not credible (i.e., case IV; move from Xj to Xk).

We will refer to the axis in Figure 2 above, spanning between incorrect and correct answers, as the COM axis. For each subject, we measured the amount of change on each trial for which confidence ratings were available from Test 2 (when misinformation was presented) to Test 3 (when misinformation was discredited). Change is measured toward the direction of the correct answer using the following equation:

$$COM = (\alpha_{Test3} * Confidence\ Test\ 3) - (\alpha_{Test2} * Confidence\ Test\ 2)$$

$$For\ correct\ answer\ \alpha = 1; for\ incorrect\ answer\ \alpha = -1$$

We found that in the Social Manipulation condition COM was significant (Fig. 3; mean COM value 36.7; significantly greater than zero $t_{(19)} = 10.8$; $p < 10^{-5}$) and significantly larger than COM values for the Credibility condition (mean COM value −0.9; $t_{(19)} = 11.2$; $p < 10^{-5}$) and the No-Information condition (mean COM value 0.51; $t_{(19)} = 12.2$; $p < 10^{-5}$). COM values tended to be larger in the Social Manipulation condition than in the Computer Manipulation condition (mean COM value 28.3; $t_{(38)} = 1.8$; trend for significance; $p = 0.08$). This is to be expected,

as misinformation delivered by humans had more of an initial influence than misinformation delivered by computers. Indeed, when controlling for the initial change in confidence that resulted from learning the opinion of others (i.e., confidence in Test 2 relative to confidence in Test 1), this effect was no longer significant; ($t_{(38)} = 1.3$; $p = 0.2$; all other comparisons remain significant $p < 10^{-5}$). Together, these behavioral results suggest that recovery from previoulsy induced errors is relatively robust, although not complete, and manifests as restoration of both judgment and confidence.

**Correcting past influence: brain mechanisms**

Functional brain imaging analysis proceeded through the following steps (detailed in the sections below) to identify a model by which past social influence is corrected. (1) First, we identified regions showing significant activation (Test 3; Fig. 1*D*) related to past presentation of unanimous false information, whether social or inanimate. Frontal and parietal regions emerged. (2) Then, we examined whether activity in any of these regions is specifically related to correction of past social influence. Here, we identified the alPFC. (3) Finally, using mediation analysis, we examined the relationship between activity in the left amygdala and left anterior hippocampus during social influence and subsequent COM activity in the left alPFC during correction. A model of correction of past social influence is then proposed.

*Identifying activity related to unanimous past misleading information*. By definition, correction of misinformation necessitates prior exposure to false information. Thus, brain regions that are important for correction should be more engaged during trials in which strong consensus of misleading information was previously present relative to trials in which it was not. To that end, we first identified brain regions where activity was greater (Test 3; Fig. 1*D*) for events in which subjects were previously presented with unanimous false information (Manipulation trials) relative to events in which information was either accurate or mixed (Credibility trials). This contrast was conducted across all subjects regardless of whether they were in the Social or Computer groups (FWE whole-brain corrected; $p < 0.05$; $k$ (cluster size) $> 50$). Significant effects were observed in five regions: The left anterior–lateral prefrontal cortex (LalPFC; peak at −34, 56, 2 (MNI); $k = 74$; Brodmann area [BA] 10), bilateral inferior parietal cortex (IPC; BA 40; 48, −56, 46; $k = 265$; −56, −58, 44; $k = 70$) the right superior medial (BA 6; 4, 34, 40; $k = 197$) and lateral prefrontal cortex (BA 9; 46, 26, 36; $k = 249$) (Fig. 4*A*; Table 1; for additional contrasts, see also Table 2).

*Identifying activity related to correction of past influence (COM) in the Social versus Computer group*. The regions above were found to be engaged during trials that were previously manipulated relative to trials that were in the Credibility condition, regardless of whether they were in the Social or Computer groups. We then asked whether activity in these regions specifically signaled a COM due to recovery from past social influence. Thus, the analysis conducted in step 1 is independent from the analysis conducted in step 2 (Kriegeskorte et al., 2009) but constrains it in a meaningful way.

COM values were entered as a parametric regressor modulating the time period from question presentation until participants' response to Manipulation questions in the Correction phase on a trial by trial basis (Test 3). For each participant, the parametric estimates averaged across all voxels, in each of the above functional ROIs, were then extracted. Unbiased statistical significance tests (Kriegeskorte et al., 2009) were performed on the comparison of the Social and Computer groups. We find that

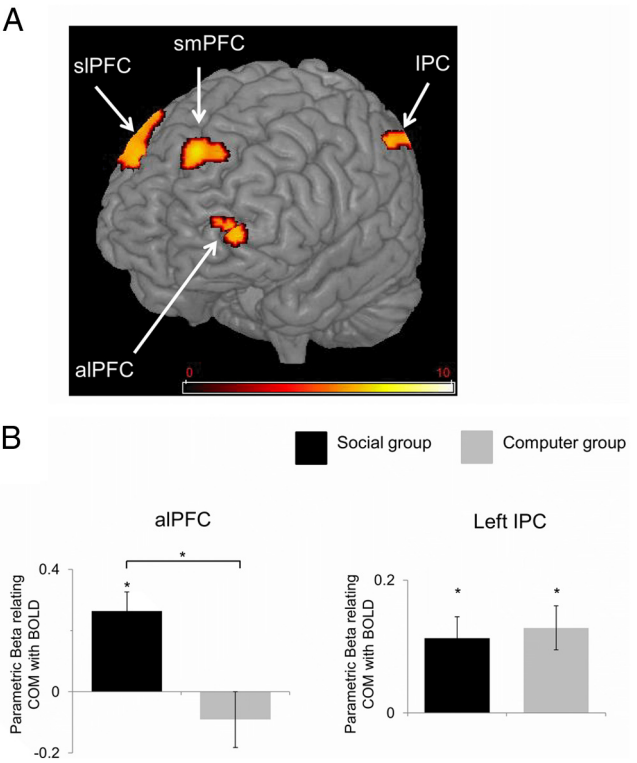


**Figure 4.** Left alPFC and left IPC are involved in the COM process (Test 3). ***A***, BOLD response during Correction phase (Test 3) was greater for trials in which consensus misleading information was previously presented (Manipulation trials) versus when it was not (Credibility trials) in the alPFC, the bilateral inferior parietal cortex, and the superior medial and lateral prefrontal cortex (whole-brain FEW-corrected: $p < 0.05$, $k > 50$ over both Social and Computer groups). ***B***, Parametric beta estimates relating COM with BOLD response in the left alPFC and left IPC for the Social group (black) and the Computer group (gray). *$p < 0.01$. Reported coordinates are in MNI space.

these parametric betas in the left alPFC were significantly greater than zero in the Social group ($t_{(19)} = 4.3$; $p < 0.001$), but not in the Computer group ($t_{(19)} = -0.8$; $p = 0.43$). Importantly, the effect of the former was significantly greater than the latter ($t_{(38)} = 2.8$; $p < 0.05$; Figure 4*B*).

Given that social influence on participants was greater than nonsocial influence, when comparing the BOLD signal related to correction, we controlled for these initial differences by adding both the magnitude of initial conformity and the change in the initial confidence after misinformation was introduced (i.e., the difference between Test 1 and Test 2 confidence) as covariates in the second-level analysis. The fMRI results remained significant; alPFC activity is more closely related to COM in the Social group than Computer group after controlling for initial influence [$F_{(3,36)} = 5.6$; $p < 0.05$)]. These results indicate that activity in the left alPFC was more likely to be related to correction of past social influence than correction of nonsocial influence and did not merely reflect the strength of initial influence.

For completeness, we also report that left IPC activation correlated with the COM in both the Social ($t_{(19)} = 3.5$; $p < 0.05$) and Computer ($t_{(19)} = 3.8$; $p < 0.05$) groups (direct comparison between groups $t_{(38)} = 0.3$; $p = 0.74$; Figure 4*B*). No other ROIs showed a significant difference between the Social and Computer groups or a significant effect for either group separately [$p > 0.3$; except for the lateral PFC ROI, which demonstrated a trend in the Social group that did not survive correction for multiple comparisons ($t_{(19)} = 2.3$; $p > 0.1$)].

**Table 1. ROI selection analysis[a]**

| Contrast | Region | MNI | t value | *p* value (after FWE whole-brain correction) | Cluster size |
|---|---|---|---|---|---|
| Manipulation > Credibility (Social and Computer groups) | Right inferior parietal cortex (BA 40) | 48, −56, 46 | 7.5 | $3 \times 10^{-4}$ FWE | 265 |
| | Superior medial prefrontal cortex (BA 8) | 4, 34, 40 | 7.1 | $7 \times 10^{-4}$ FWE | 197 |
| | alPFC (BA 10) | −34, 56, 2 | 6.8 | 0.002 FWE | 74 |
| | Superior lateral prefrontal cortex (BA 9) | 46, 26, 36 | 6.7 | 0.003 FWE | 249 |
| | Left inferior parietal cortex (BA 40) | −56, −58, 44 | 6.6 | 0.003 FWE | 72 |

[a]Regions more active for Manipulation versus Credibility regardless of group. Minimum cluster size 50.

**Table 2. Manipulation versus Credibility for Social and Computer groups separately[a,b]**

| Contrast | Region | MNI | t value | *p* value (cluster extent) | Cluster size |
|---|---|---|---|---|---|
| Manipulation > Credibility (Social group only) | IPC (BA 40, BA 39) | 48, −56, 46; | 6.2 | $8 \times 10^{-6}$ | 1010 |
| | | −56, −58, 44 | 5.3 | $9 \times 10^{-5}$ | 680 |
| | | 14, −54, 34 | 6.1 | 0.008 | 274 |
| | Inferior frontal gyrus (BA 47, BA 45) | 30, 24, −10 | 5.2 | $4 \times 10^{-5}$ | 781 |
| | | 56, 24, 2 | 5.0 | 0.02 | 194 |
| | Lateral prefrontal cortex (BA 6, BA 9) | 46, 24, 36 | 6.0 | $2 \times 10^{-5}$ | 876 |
| | | −40, 14, 56 | 4.7 | 0.01 | 254 |
| | alPFC (BA 10) | −36, 56, 2 | 5.9 | 0.002 | 399 |
| | Medial frontal cortex (BA 8) | 6, 34, 40 | 5.4 | $6 \times 10^{-7}$ | 1396 |
| Manipulation > credibility (Computer group only) | Occipital cortex, cuneus (BA 17) | −12, −94, −2 | 5.2 | 0.03 | 262 |
| | Medial frontal cortex (BA 8) | 4, 34, 42 | 4.8 | 0.0001 | 795 |
| | IPC (BA 40) | 42, −50, 44 | 4.5 | 0.03 | 245 |
| | alPFC (BA 10) | −30, 54, 2 | 4.2 | 0.04 | 210 |
| Manipulation > Credibility (Social > Computer; or Social < Computer) | No significant clusters | | | | |

[a]FDR cluster extent correction <0.05, cluster defining threshold 0.001 uncorrected.

[b]Testing the additional regions found in Table 2 (excluding the alPFC and IPS) for COM did not result in significant correlations ($p > 0.2$).

It is of note that the more refined COM model fitted the brain data better than a standard contrast approach, which compares trials in which subjects successfully corrected past social influence versus trials in which they failed to do so. Specifically, the COM model $\beta$ values were significantly greater than the simple contrast model betas for left alPFC ($t_{(19)} = 2.6$; $p < 0.05$).

These results suggest that activity in the left alPFC is related to recovery from social influence. We next characterized the brain's processing, going from activity during initial social influence to subsequent correction via the left alPFC.

*Characterizing a model for correction of past social influence (COM).* Correction of past social influence is likely dependent not only on processes that occur after low credibility is revealed but also on processes that occurred when the erroneous influence was introduced in the first place. We previously reported that heightened (left) amygdala-anterior hippocampal connectivity during social manipulation (Manipulation phase, Test 2; Fig. 1*C*) predicted a long-lasting effect of the erroneous social information on the subjects' memory (i.e., long-term errors in memory) (Edelson et al., 2011). We thus hypothesized that amygdala activation during the time of social influence may alter memory representations, rendering subsequent engagement of left alPFC correction mechanisms less likely. In this case, amygdala influence on COM would involve a two-step process. First, during the time of social influence, amygdala activity would affect hippocampal mnemonic representations. These changes will in turn be related to subsequent left alPFC mediation of COM during the Correction phase. To test this hypothesis, we conducted a mediation analysis (Atlas et al., 2010; Hayes, 2013, Wager et al., 2008), following the steps of modern revised mediation approaches, that included activation in left amygdala (Test 2), left anterior hippocampus (Test 2), and left alPFC (Test 3) as predictors of COM. We first extracted left amygdala and left anterior hippocampal activity (averaged over all voxels in these anatomically defined regions) during the time subjects were first exposed to false information (i.e., Test 2, Fig. 1*C*) for each specific trial. We then performed the same calculation for the left alPFC region during recovery (Test 3). Next, using linear regression models, we calculated the weights for each path per subject (Fig. 5).

Left amygdala activity during initial exposure to the opinion of others was inversely related to the subjects' subsequent tendency to change their minds when it was revealed that those opinions were fabricated [referred to in mediation analysis terminology (Hayes, 2013) as the Total effect = −11.4; $z(19) = 2.1$; $p < 0.05$; all parameter estimates are unstandardized]. Importantly, this relationship was at least partially mediated by an indirect effect of the left anterior hippocampus during exposure to the opinions of others and alPFC activity during the time of correction. (parameter estimate of indirect effect of interest = −2.5; $z(19) = -2.4$; $p < 0.05$; adding the indirect pathway significantly reduced the direct effect of the amygdala on COM parameter estimate = −5.3; $z(19) = -0.8$; $p = 0.46$; Sobel test $p < 0.05$). The model showed that left amygdala activity correlated with left hippocampal activity during exposure to social influence (parameter estimate = 0.53; $z(19) = 3.9$; $p < 0.05$) which was inversely related to left alPFC activation during correction (parameter estimate = −0.44; $z(19) = -2.8$; $p < 0.05$), which was in turn significantly related to COM (parameter estimate = 4.5; $z$

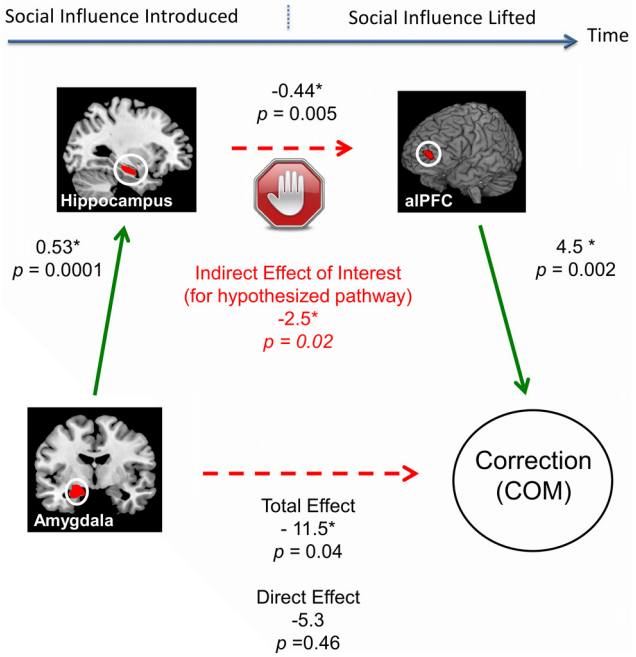


**Figure 5.** Sequential model for correction of past social influence. Mediation results suggest that amygdala-related restriction of correction is mediated via hippocampal and left alPFC activity (indirect pathway of interest). The arrows indicate that the separate components of the indirect pathway are significant. After taking into account all indirect effects, the direct effect of the amygdala on COM is no longer significant. The total effect represents the summation of all possible indirect and direct pathways. Values represent unstandardized $\beta$ parameter estimates for each path. *$p < 0.05$. Dashed line indicates correlation over different time periods and thus may be mediated by additional factors. Red and green arrows indicate a negative and positive correlation, respectively.

(19) = 3.0; $p < 0.05$). Adding the hippocampus as a predictor in the regression model significantly increased the explained variance in COM scores ($R^2$ adjusted for model complexity = 0.19 vs 0.24; $t_{(19)} = 3.8$; $p < 0.05$), indicating that the hippocampus played a role in the indirect pathway. These results indicate that the indirect pathway of interest explains a significant part of the relationship between amygdala activity in Test 2 and COM in Test 3. For completeness, we conducted the same mediation model in the Computer group. The mediation was not significant in this group ($p > 0.4$). This is to be expected as activity neither in the amygdala nor in the alPFC was related to COM in the Computer group.

*Testing other models.* For completeness, we tested whether alPFC activation in the mediation model could be substituted with activation of the other ROIs identified in Figure 4 (smPFC, dlPFC, and IPS). We did not find this to be the case, as the indirect effects were nonsignificant ($p > 0.4$), even for the left IPS in which activity correlated with COM in Test 3.

We also tested whether medial temporal lobe activation during the time of social influence correction (i.e., Test 3 rather than Test 2) mediated COM via alPFC activity. Using the left amygdala and anterior hippocampus activation values from the Correction phase (Test 3 rather than Test 2 as in the first model) did not yield a significant effect between amygdala activity in Test 3 and COM values or a significant mediation ($p > 0.3$). This result suggests that socially induced mnemonic alterations, mediated by the medial temporal lobe, may have already occurred at the stage of initial exposure to social influence rather than in the Correction phase. Thus, our results do not imply or necessitate a direct link between medial temporal lobe and alPFC. We also note that other regions may play a role in the process. Our model does not necessarily suggest that the regions identified are the sole regions involved; rather, we suggest a function for these regions in this process.

## Discussion

The human social environment is dynamic and mandates flexible mechanisms that enable us both to learn from others and to reverse such learning when that influence is no longer valid (Humphrey, 1976; Byrne and Whiten, 1989; Adolphs, 1999; Dunbar, 2003; Olsson and Phelps, 2007; Campbell-Meiklejohn et al., 2010; Zaki et al., 2011; Engelmann and Hein, 2013; Haun et al., 2013). The current study describes a brain model underlying this ability.

We first demonstrate that left alPFC activity reflected whether the evidence shift was sufficient to induce an adjustment in judgment. Our mediation analysis suggests an interaction between socially induced false memory encoding and the subsequent ability to change one's mind. We then describe how this complex process may unfold over two time points (Fig. 5). Specifically, we suggest that amygdala activity during the time of initial social influence affects the hippocampal-dependent memory system, presumably altering mnemonic representations, which in turn are related to restricted subsequent prefrontal correction mechanisms when influence is lifted. It is possible that amygdala activity leads to strongly encoded false memories that dominate the original representations. This in turn may restrict the possibility of recovery. These findings illuminate the process by which errors are, or fail to be, corrected and highlight how social influence restricts subsequent correction, even when that influence is later discredited.

We found that alPFC activity was more tightly correlated with correction of past social influence than nonsocial influence, even after controlling for the initial size of influence. Our results tie together past findings in primates and humans demonstrating that the alPFC (sometimes referred to as rostrolateral PFC, lateral orbitofrontal cortex, frontopolar cortex, or ventrolateral PFC, all with similar spatial coordinates, e.g., Clark et al., 2004; Mendelsohn et al., 2008; Bunge et al., 2009; Boorman et al., 2011; Sakaki et al., 2011; Badre et al., 2012) possesses the necessary capabilities to partake of restorative processes and represent the value of switching to a counterfactual or alternative choice (Cabeza and Nyberg, 1997; Ramnani and Owen, 2004; Koechlin and Hyafil, 2007; Mendelsohn et al., 2008; Bunge et al., 2009; Rushworth et al., 2011; Sakaki et al., 2011; Badre et al., 2012).

Our findings suggest that the alPFC tracks a subjective change of mind and may improve selection and updating of competing items by monitoring the change in evidence supporting the individuals' initially selected option versus the option endorsed by the group. Exceeding a threshold of alPFC activation may be related to a change in behavioral judgment subserving successful recovery from past influence. The function of alPFC is especially important when environmental demands rapidly change and mandate an adjustment of behavior (Rogers et al., 2000; Kringelbach and Rolls, 2003; Clark et al., 2004; Morris and Dolan, 2004; Elliott and Deakin, 2005; Remijnse et al., 2005; Badre et al., 2012).

The alPFC is involved in general control mechanisms (Christoff and Gabrieli, 2000; Ramnani and Owen, 2004; Elliott and Deakin, 2005; Gilbert et al., 2006; Koechlin and Hyafil, 2007; Bunge et al., 2009; Badre et al., 2012; Ruff et al., 2013), including mnemonic control (Lepage et al., 2000; Mendelsohn et al., 2008; Sakaki et al., 2011), such as reality (Johnson and Raye, 1981;

Johnson et al., 1993) and source monitoring (Thompson-Schill et al., 1997; Wagner et al., 2004), as well as the ability to evaluate one's own cognitive operations (Fleming et al., 2010, 2012) and to process internally versus externally generated information (Christoff and Gabrieli, 2000; Gilbert et al., 2006). Thus, the general function of this region may be preferentially engaged depending on social contextual factors (Kringelbach and Rolls, 2003; David et al., 2006; Spitzer et al., 2007; Mitchell et al., 2009; Campbell-Meiklejohn et al., 2010; Raposo et al., 2011; Boorman et al., 2013; Ruff et al., 2013).

Challenges in the social environment are assumed to exert important selective pressures in the evolution of the hominid brain, particularly frontal-dependent faculties (Adolphs, 1999; Dunbar, 2003; Haun et al., 2013). The social environment can be highly dynamic, mandating that individuals not retain overly rigid representations and beliefs. The ability to change one's mind when the social environment changes is probably one of the crucial processes for our survival (Humphrey, 1976; Byrne and Whiten, 1989; Elliott and Deakin, 2005). Thus, studying recovery from social influence may present an advantageous way of studying recovery mechanisms from powerful past influences. The current study demonstrates how the human brain may achieve this flexibility. Whereas long-term distortion resulting from the influence of others is mediated by medial–temporal activity (Edelson et al., 2011; Deuker et al., 2013), we show here that reversing these effects to recover an original veridical belief is mediated by prefrontal activity. Importantly, we suggest that the final outcome is related to both systems, as restoration abilities in the alPFC may be restricted by past amygdala's modulation of hippocampal-dependent memory system. The evolution of these mechanisms may have helped humans and other social animals by counterbalancing an adaptive tendency to conform with a useful degree of plasticity within the social milieu.

## References


Adolphs R (1999) Social cognition and the human brain. Trends Cogn Sci 3:469–479. CrossRef Medline

Allen V (1975) Social support for nonconformity. Adv Exp Soc Psychol 8:1–43. CrossRef

Asch SE (1952) Group forces in the modification and distortion of judgments. In: Social psychology, pp 450–501. Englewood Cliffs, NJ: Prentice-Hall.

Atlas LY, Bolger N, Lindquist MA, Wager TD (2010) Brain mediators of predictive cue effects on perceived pain. J Neurosci 30:12964–12977. CrossRef Medline

Badre D, Doll BB, Long NM, Frank MJ (2012) Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73:595–607. CrossRef Medline

Berns GS, Chappelow J, Zink CF, Pagnoni G, Martin-Skurski ME, Richards J (2005) Neurobiological correlates of social conformity and independence during mental rotation. Biol Psychiatry 58:245–253. CrossRef Medline

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev 113:700–765. CrossRef Medline

Bonnici H, Chadwick M, Kumaran D, Hassabis D, Weiskopf N, Maguire EA (2012) Multi-voxel pattern analysis in human hippocampal subfields. Front Hum Neurosci 6.

Boorman ED, Behrens TE, Rushworth MF (2011) Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. PLoS Biol 9.

Boorman ED, O'Doherty JP, Adolphs R, Rangel A (2013) The behavioral and neural mechanisms underlying the tracking of expertise. Neuron 80:1558–1571. CrossRef Medline

Braun K, Loftus E (1998) Advertising's misinformation effect. Appl Cogn Psychol 12:569–591. CrossRef

Bunge SA, Helskog EH, Wendelken C (2009) Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. Neuroimage 46:338–342. CrossRef Medline

Byrne R, Whiten A (1989) Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans (Byrne R, Whiten A, eds). Oxford: Oxford UP.

Cabeza R, Nyberg L (1997) Imaging cognition: an empirical review of PET studies with normal subjects. J Cogn Neurosci 9:1–26. CrossRef Medline

Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith CD (2010) How the opinion of others affects our valuation of objects. Curr Biol 20:1165–1170. CrossRef Medline

Chadwick MJ, Bonnici HM, Maguire EA (2012) Decoding information in the human hippocampus: a user's guide. Neuropsychologia 50:3107–3121. CrossRef Medline

Chambers KL, Zaragoza MS (2001) Intended and unintended effects of explicit warnings on eyewitness suggestibility: evidence from source identification tests. Mem Cogn 29:1120–1129. CrossRef Medline

Charpentier CJ, Moutsiana C, Garrett N, Sharot T (2014) The brain's temporal dynamics from a collective decision to individual action. J Neurosci 34:5816–5823. CrossRef Medline

Christoff K, Gabrieli JDE (2000) The frontopolar cortex and human cognition: evidence for a rostrocaudal hierarchical organization within the human prefrontal cortex. Psychobiology 28:168–186.

Cisler JM, Bush K, Steele JS (2014) A comparison of statistical methods for detecting context-modulated functional connectivity in fMRI. Neuroimage 84:1042–1052. CrossRef Medline

Clark L, Cools R, Robbins TW (2004) The neuropsychology of ventral prefrontal cortex: decision-making and reversal learning. Brain Cogn 55:41–53. CrossRef Medline

David N, Bewernick BH, Cohen MX, Newen A, Lux S, Fink GR, Shah NJ, Vogeley K (2006) Neural representations of self versus other: visual-spatial perspective taking and agency in a virtual ball-tossing game. J Cogn Neurosci 18:898–910. CrossRef Medline

De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. Nat Neurosci 16:105–110. CrossRef Medline

Deuker L, Müller AR, Montag C, Markett S, Reuter M, Fell J, Trautner P, Axmacher N (2013) Playing nice: a multi-methodological study on the effects of social conformity on memory. Front Hum Neurosci 7:79. CrossRef Medline

Deutsch M, Gerard HB (1955) A study of normative and informational influences upon individual judgement. J Abnorm Psychol 51:629–636. Medline

Dunbar RIM (2003) The social brain: mind, language, and society in evolutionary perspective. Annu Rev Anthropol 32:163–181. CrossRef

Echterhoff G, Hirst W, Hussy W (2005) How eyewitnesses resist misinformation: social postwarnings and the monitoring of memory characteristics. Mem Cogn 33:770–782. CrossRef Medline

Edelson M, Sharot T, Dolan RJ, Dudai Y (2011) Following the crowd: brain substrates of long-term memory conformity. Science 333:108–111. CrossRef Medline

Elliott R, Deakin B (2005) Role of the orbitofrontal cortex in reinforcement processing and inhibitory control: evidence from functional magnetic resonance imaging studies in healthy human subjects. Int Rev Neurobiol 65:89–116. CrossRef Medline

Engelmann JB, Hein G (2013) Contextual and social influences on valuation and choice. Prog Brain Res 202:215–237. CrossRef Medline

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. Science 329:1541–1543. CrossRef Medline

Fleming SM, Huijgen J, Dolan RJ (2012) Prefrontal contributions to metacognition in perceptual decision making. J Neurosci 32:6117–6125. CrossRef Medline

Gilbert SJ, Spengler S, Simons JS, Steele JD, Lawrie SM, Frith CD, Burgess PW (2006) Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. J Cogn Neurosci 18:932–948. CrossRef Medline

Haun DB, van Leeuwen EJ, Edelson MG (2013) Majority influence in children and other animals. Dev Cogn Neurosci 3:61–71. CrossRef Medline

Hayes A (2013) Introduction to mediation, moderation, and conditional process analysis: a regression-based approach. New York: Guilford.

Hirst W, Echterhoff G (2012) Remembering in conversations: the social sharing and reshaping of memories. Annu Rev Psychol 63:55–79. CrossRef Medline

Howell DC (1997) Statistical methods for psychology, Ed 4. London: Duxbury.
Humphrey N (1976) The social function of intellect. In: Growing points in ethology (Bateson PPG, Hinde RA, eds), pp 303–317. Cambridge: Cambridge UP.
Johnson M, Raye C (1981) Reality monitoring. Psychol Rev 88:67–85. CrossRef
Johnson M, Hashtroudi S, Lindsay D (1993) Source monitoring. Psychol Bull 114:3–28. CrossRef Medline
Klucharev V, Hytönen K, Rijpkema M, Smidts A, Fernández G (2009) Reinforcement learning signal predicts social conformity. Neuron 61:140–151. CrossRef Medline
Koechlin E, Hyafil A (2007) Anterior prefrontal function and the limits of human decision-making. Science 318:594–598. CrossRef Medline
Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540. CrossRef Medline
Kringelbach ML, Rolls ET (2003) Neural correlates of rapid reversal learning in a simple model of human social interaction. Neuroimage 20:1371–1383. CrossRef Medline
Lancaster JL, Rainey LH, Summerlin JL, Freitas CS, Fox PT, Evans AC, Toga AW, Mazziotta JC (1997) Automated labeling of the human brain: a preliminary report on the development and evaluation of a forward-transform method. Hum Brain Mapp 5:238–242. CrossRef Medline
Lepage M, Ghaffar O, Nyberg L, Tulving E (2000) Prefrontal cortex and episodic memory retrieval mode. Proc Natl Acad Sci U S A 97:506–511. CrossRef Medline
Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. Psychol Sci 13:106–131. Medline
Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239. CrossRef Medline
Meade ML, Roediger HL (2002) Explorations in the social contagion of memory. Mem Cogn 30:995–1009. CrossRef Medline
Mendelsohn A, Chalamish Y, Solomonovich A, Dudai Y (2008) Mesmerizing memories: brain substrates of episodic memory suppression in post-hypnotic amnesia. Neuron 57:159–170. CrossRef Medline
Mitchell JP, Ames DL, Jenkins AC, Banaji MR (2009) Neural correlates of stereotype application. J Cogn Neurosci 21:594–604. CrossRef Medline
Morris JS, Dolan RJ (2004) Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. Neuroimage 22:372–380. CrossRef Medline
Mumford JA, Turner BO, Ashby FG, Poldrack RA (2012) Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59:2636–2643. CrossRef Medline
Olsson A, Phelps EA (2007) Social learning of fear. Nat Neurosci 10:1095–1102. CrossRef Medline
Ortmann A, Hertwig R (2002) The costs of deception: evidence from psychology. Exp Econ 5:111–131. CrossRef
Ramnani N, Owen AM (2004) Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. Nat Rev Neurosci 5:184–194. CrossRef Medline
Raposo A, Vicens L, Clithero JA, Dobbins IG, Huettel SA (2011) Contributions of frontopolar cortex to judgments about self, others and relations. Soc Cogn Affect Neurosci 6:260–269. CrossRef Medline
Remijnse PL, Nielen MM, Uylings HB, Veltman DJ (2005) Neural correlates of a reversal learning task with an affectively neutral baseline: an event-related fMRI study. Neuroimage 26:609–618. CrossRef Medline
Resulaj A, Kiani R, Wolpert DM, Shadlen MN (2009) Changes of mind in decision-making. Nature 461:263–266. CrossRef Medline
Rissman J, Gazzaley A, D'Esposito M (2004) Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage 23:752–763. CrossRef Medline
Rogers RD, Andrews TC, Grasby PM, Brooks DJ, Robbins TW (2000) Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. J Cogn Neurosci 12:142–162. CrossRef Medline
Ross L, Lepper MR, Hubbard M (1975) Perseverance in self-perception and social perception. J Pers Soc Psychol 32:880–892. CrossRef Medline
Ruff CC, Ugazio G, Fehr E (2013) Changing social norm compliance with noninvasive brain stimulation. Science 342:482–484. CrossRef Medline
Rushworth MF, Noonan MP, Boorman ED, Walton ME, Behrens TE (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70:1054–1069. CrossRef Medline
Sakaki M, Niki K, Mather M (2011) Updating existing emotional memories involves the frontopolar/orbito-frontal cortex in ways that acquiring new emotional memories does not. J Cogn Neurosci 23:3498–3514. CrossRef Medline
Schacter DL, Loftus EF (2013) Memory and law: what can cognitive neuroscience contribute? Nat Neurosci 16:119–123. CrossRef Medline
Schacter DL, Guerin SA, St Jacques PL (2011) Memory distortion: an adaptive perspective. Trends Cogn Sci 15:467–474. CrossRef Medline
Schiller D, Levy I, Niv Y, Ledoux J, Phelps E (2008) From fear to safety and back: reversal of fear in the human brain. J Neurosci 5:11517–11525. CrossRef Medline
Sherif M (1936) The psychology of social norms. Oxford: Oxford UP.
Skurnik I, Yoon C, Park D, Schwarz N (2005) How warnings about false claims become recommendations. J Consum Res 31:713–724. CrossRef
Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E (2007) The neural signature of social norm compliance. Neuron 56:185–196. CrossRef Medline
Stang DDJ (1976) Ineffective deception in conformity research: some causes and consequences. Eur J Soc Psychol 6:353–367. CrossRef
Surowiecki J (2004) The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. New York: Doubleday.
Thompson-Schill SL, D'Esposito M, Aguirre GK, Farah MJ (1997) Role of left inferior prefrontal cortex in retrieval of semantic knowledge: a reevaluation. Proc Natl Acad Sci U S A 94:14792–14797. CrossRef Medline
Vickers D (1970) Evidence for an accumulator model of psychophysical discrimination. Ergonomics 13:37–58. CrossRef Medline
Vlaev I, Chater N, Stewart N, Brown GD (2011) Does the brain calculate value? Trends Cogn Sci 15:546–554. CrossRef Medline
Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN (2008) Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron 59:1037–1050. CrossRef Medline
Wagner A, Bunge S, Badre D (2004) Cognitive control, semantic memory, and priming: contributions from prefrontal cortex. In: The cognitive neurosciences, Ed 3 (Gazzaniga MS, ed), pp 709–725. Cambridge, MA: Massachusetts Institute of Technology.
Wright DB, Memon A, Skagerberg EM, Gabbert F (2009) When eyewitnesses talk. Curr Dir Psychol Sci 18:174–178. CrossRef
Zaki J, Schirmer J, Mitchell JP (2011) Social influence modulates the neural computation of value. Psychol Sci 22:894–900. CrossRef Medline