Following the Crowd: Brain Substrates of Long-Term Memory Conformity

Micah Edelson, *et al.*

Science **333**, 108 (2011);
DOI: 10.1126/science.1203557

Following the Crowd: Brain Substrates of Long-Term Memory Conformity

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

Human memory is strikingly susceptible to social influences, yet we know little about the underlying mechanisms. We examined how socially induced memory errors are generated in the brain by studying the memory of individuals exposed to recollections of others. Participants exhibited a strong tendency to conform to erroneous recollections of the group, producing both long-lasting and temporary errors, even when their initial memory was strong and accurate. Functional brain imaging revealed that social influence modified the neuronal representation of memory. Specifically, a particular brain signature of enhanced amygdala activity and enhanced amygdala-hippocampus connectivity predicted long-lasting but not temporary memory alterations. Our findings reveal how social manipulation can alter memory and extend the known functions of the amygdala to encompass socially mediated memory distortions.

Our memories are often inaccurate. Ubiquitous sources of false recollection are social pressure and interpersonal influence (*1*–*4*). This phenomenon, dubbed "memory conformity" (*4*), is encountered in a variety of contexts, including social interactions, mass media exposure, and eyewitness testimony. In such settings an individual may change veridical recollections of past events to match a false account provided by others (*1*–*6*). Although these social influences on memory have been extensively demonstrated (*1*–*5*), the underlying neurobiology of this process is unknown.

Conformity may present in two forms, which initially convey similar explicit behavior but are fundamentally different (*7*, *8*). In one type, known as private conformity, an individual's recollection may genuinely be altered by social influence, resulting in long-lasting, persistent memory errors (*1*, *4*, *5*, *7*). In such circumstances, even when social influence is removed, the individuals will persist in claiming an erroneous memory as part of their own experience (*7*, *9*). Private conformity could hence be considered a bona fide memory change. In the second type, known as public conformity, individuals may choose to outwardly comply, providing an account that fits that of others, but inwardly maintain certitude in their own original memory. Public conformity can be dispelled when the veracity of the socially transferred information abates (*7*, *10*, *11*). Thus, errors induced by public conformity are transient (*7*, *9*) and appear to represent a change in behavior in the absence of lasting alterations to a memory engram.

Although private and public memory conformity are often behaviorally indistinguishable, they reflect different cognitive processes (*7*, *8*). These processes are probably mediated by distinct activation in interconnected brain circuits previously found to be active in mnemonic functions and social cognition (such as the hippocampal complex, amygdala, and frontal regions) (*12*–*18*). Here, we set out to characterize the brain mechanisms that lead to both types of conformity.

Our experimental protocol included four phases spanning a 2-week period (Fig. 1A). Thirty adult participants (12 females, age 28.6 ± 0.8, mean ± SEM) viewed an eyewitness-style documentary on a large screen in groups of five. Three days after viewing, participants returned to the lab individually and completed a memory test (test 1). Test 1 served to assess the participants' baseline accuracy and confidence before the manipulation stage. Four days later, participants returned to the lab and answered the same memory questions while being scanned with functional magnetic resonance imaging (fMRI) (test 2). On this occasion, a manipulation was introduced in an attempt to induce conformity.

Before responding during this test, participants were presented with answers they were led to believe were given by their four fellow co-observers, whose photographs were provided with their corresponding answers (Fig. 1A). In a subset of trials, for which the target participant originally had a confident veridical memory (as identified by test 1), the answers provided by the four co-observes were all false (manipulation condition, 80 questions). In matched control trials, the letter X was presented instead of the co-observers' answers (no-manipulation condition, 25 questions). Pilot data indicated that the use of manipulation and no-manipulation conditions alone would raise suspicion in the participants' minds that the answers given by the co-observers were fabricated. Therefore we added credibility trials in which different patterns of co-observer answers were provided (Fig. 1B).

One week later, the participants returned to the lab and were informed that the answers given by the co-observers during the previous fMRI session were in fact determined randomly. This rendered the socially conveyed information previously provided as uninformative. The participants were then requested to complete the memory test again (test 3) based on their original memory of the movie. Finally, the participants were debriefed. Participants with excessive head movements in the scanner or suspected brain pathology and those that indicated suspicion of the manipulation were excluded from the analysis, resulting in a final number of participants (N) = 20.

Our behavioral data revealed that our manipulation induced memory errors (Fig. 2A). Strikingly, participants conformed to the majority opinion in 68.3 ± 2.9% of manipulation trials, giving a false answer to questions they had previously answered correctly with relatively high confidence. This was not due to forgetting, because in the no-manipulation condition, incorrect answers were given in only 15.5 ± 1.7% of the questions [Student's t test (df 19) = 16.9, $P < 10^{-7}$]. When social influence was removed (test 3), participants reverted to their original correct answer in 59.2 ± 2.3% of the previously conformed trials (transient errors) but maintained erroneous answers in 40.8% (persistent errors). Confidence ratings in persistent and transient errors did not differ either before or after the manipulation stage (Fig. 2B). During the manipulation stage, confidence ratings in transient errors were significantly lower than in persistent errors [t (19) = 6.9, $P < 10^{-5}$]. Differences in confidence levels were controlled for in the fMRI analysis by means of a covariate [supporting online material (SOM)].

Our brain imaging data indicated that at the time of exposure to social influence, distinct brain signatures characterized instances of memory conformity that would result in persistent and transient errors. We first performed analysis on a priori anatomically defined regions of interest (ROIs) selected by virtue of being widely implicated in memory encoding and maintenance (the bilateral anterior hippocampus, bilateral posterior hippocampus, and bilateral parahippocampal gyrus) and in social-emotional processing (bilateral amygdala) (*12*–*25*). Brain activity was averaged across all voxels in each ROI for the three conditions of interest (persistent errors, transient errors, and instances when participants did not conform to the erroneous information; i.e., nonconformity). In all regions, except for the left posterior hippocampus, the blood oxygen level–dependent (BOLD) signal was greater during trials that subsequently resulted in persistent memory errors relative to trials that resulted in transient errors or nonconformity (Fig. 3A). No significant difference was found between transient error and nonconformity trials in these regions.

To examine whether other brain regions differentiate between persistent and transient errors, we conducted a whole-brain exploratory analysis. Greater activity during trials resulting in persistent errors versus trials resulting in transient errors was found in four regions, all in the medial temporal lobe (MTL, Fig. 3B): the left amygdala (–22,–8,–10), right hippocampus (28,–22,–12), right parahippocampal gyrus (PHG, 36,–48,–10), and a region bordering the left PHG and occipital

[1]Department of Neurobiology, Weizmann Institute of Science, Israel. [2]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK.

*To whom correspondence should be addressed. E-mail: micah.edelson@weizmann.ac.il

cortex (–22,–54,–10), [$P < 0.001$, cluster threshold (k) > 10]. In the opposite comparison (transient versus persistent errors), enhanced activation was found in the bilateral dorsal anterior cingulate cortex (ACC, Brodmann area 32; –12,22,42; 8,20,46).

A striking activation in both aforementioned analyses was found in the left amygdala. A behavioral control study (SOM) indicated that elevated activation in the amygdala during trials that resulted in persistent errors was not due to heightened emotional arousal during these trials. Nor were these errors related to questions associated with greater emotional content. Rather, heightened amygdala activation seemed specific to socially induced memory change.

The amygdala plays a key role in social and emotional processing and modulates memory-related hippocampal activity (*13*–*23*). It is strategically placed for this function, having rich anatomical connections with the hippocampal complex (the anterior hippocampus in particular) as well as with neocortical areas (*13*–*16*, *23*, *26*). The amygdala is thus a prime candidate for mediating social effects on memory, most likely involving its interactions with other brain regions (*13*, *14*). This consideration motivated us to carry out a functional connectivity analysis, using a psychophysiological interaction (PPI) approach (*27*). This analysis showed heightened functional connectivity between the left amygdala and bilateral anterior hippocampus within anatomically defined ROIs, during trials that subsequently resulted in persistent memory errors as opposed to transient errors and nonconformity (Fig. 4A).

We also sought to identify which brain regions responded to the information presented by the co-observers (SOM). To this end, trials in which misleading information was presented

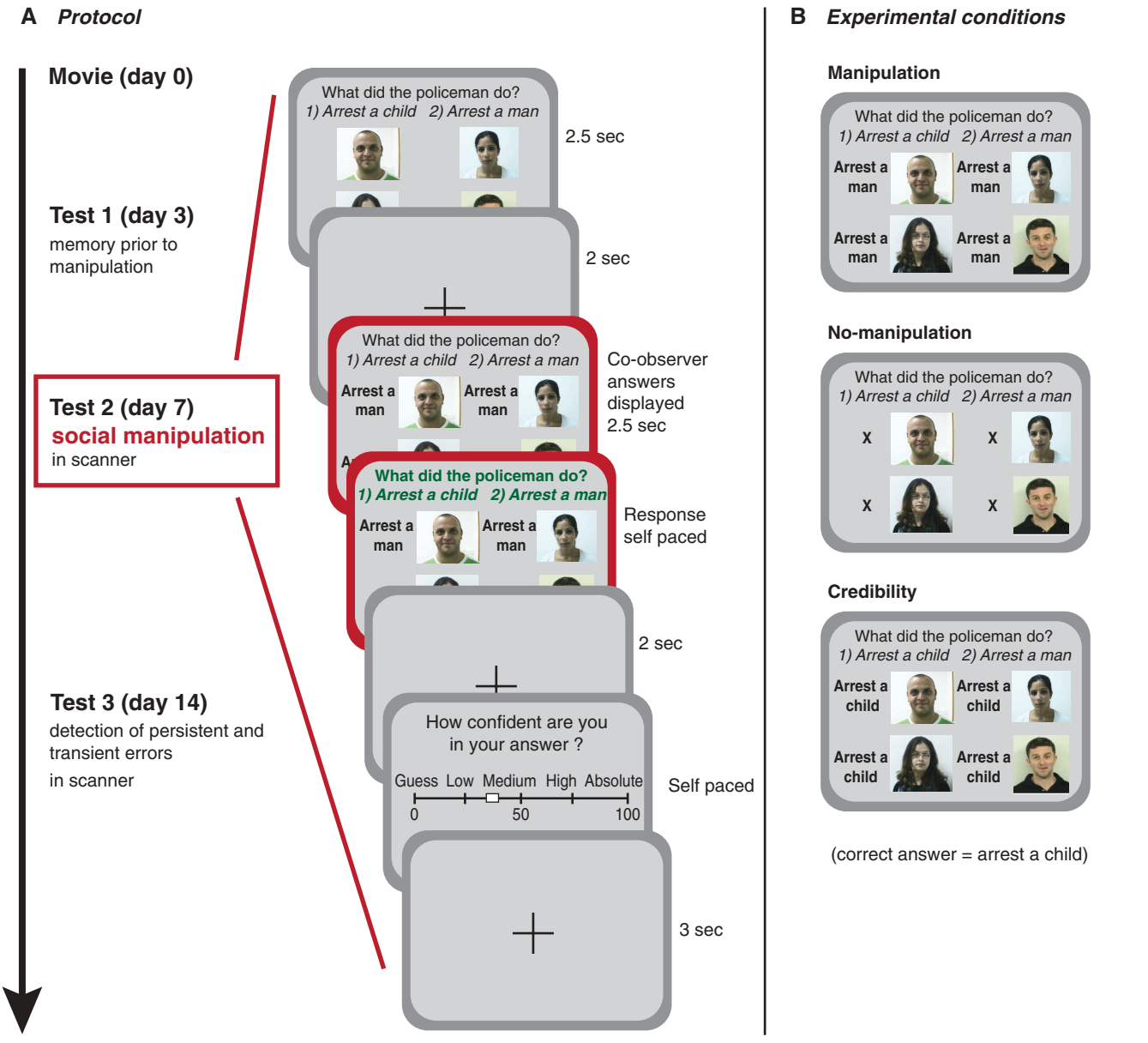

Fig. 1. Experimental outline. (**A**) Participants viewed the movie in groups of five and subsequently performed three memory tests individually. Test 1 served to assess the participants' initial memory and confidence before the social manipulation administered in test 2. Test 3 served to identify memory errors that persisted after the social manipulation was removed. For the test 2 scanning session, the question and possible answers were presented for 2.5 s, followed by the fabricated co-observers' answers for 2.5 s. Subsequently, a font color change indicated that the participants were allowed to respond. Finally, confidence ratings were provided. (**B**) Illustration of the different experimental conditions: the manipulation condition in which all co-observers' answers were incorrect, the no-manipulation condition in which the letter X was displayed instead of co-observers' answers, and the credibility condition in which variable patterns of co-observers' answers were displayed (SOM).

(the manipulation condition) were contrasted with the no-manipulation condition. Five regions (fig. S1A) were identified in the frontal and occipital cortex. Further analysis of brain activity in these regions (fig. S1B) suggests that they are involved in non-mnemonic processes, such as conflict monitoring (*28*–*31*) in the face of competing memories (*32*–*34*).

Were our findings driven merely by the presentation of additional information regardless of social context? To answer this question, we performed a control fMRI experiment using a non-social medium to convey misinformation (SOM). Participants underwent a similar protocol to that of our main experiment. However, in memory test 2, 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 (*30*). Conformity in this case was significantly lower (45.3 ± 4.7%) than in the social manipulation described earlier (68.3 ± 2.9%) but significantly higher than with no manipulation at all (15.0 ± 2.4%) [t (38) = 4.2 and t (19) = –5.7, respectively; $P < 0.0002$)].

Analysis of BOLD signal in the a priori MTL ROIs revealed an interaction between memory (persistent errors and transient errors) and experimental manipulation (social and nonsocial) in the bilateral amygdala ($P < 0.05$). This interaction was driven by greater activation in trials resulting in persistent memory errors relative to transient errors in the social manipulation, but not in the nonsocial manipulation (Fig. 3A). These results suggest that enhanced activity in these regions is related specifically to socially induced persistent memory errors. In contrast, the right anterior and posterior hippocampus and left PHG revealed a main effect of memory ($P < 0.05$), where there was greater activity during trials resulting in persistent relative to transient errors regardless of manipulation type ($P < 0.05$) (Fig. 3A). Thus, the BOLD signal in these regions was associated with long-lasting memory errors irrespective of the medium by which information was conveyed. Results of a functional connectivity analysis between the left amygdala and bilateral anterior hippocampus showed a significant interaction ($P < 0.05$). Heightened connectivity was seen during trials that resulted in persistent errors relative to transient errors, a pattern specific to the social manipulation (Fig. 4B). Our control experiment's results hence indicate that heightened amygdala activation and enhanced connectivity with the hippocampus are specific to socially induced memory changes, whereas hippocampal complex activation differentiates between persistent and transient errors regardless of the source of influence.

Our results indicate that memory is highly susceptible to alteration due to social influence, creating both transient and persistent errors. After over a century of intensive behavioral research into social influences on memory (*35*), this study now provides a brain account of this phenomenon. Our findings suggest a mechanism by which social influence produces long-lasting alterations in memory, and they highlight the critical role of the amygdala in mediating this influence.

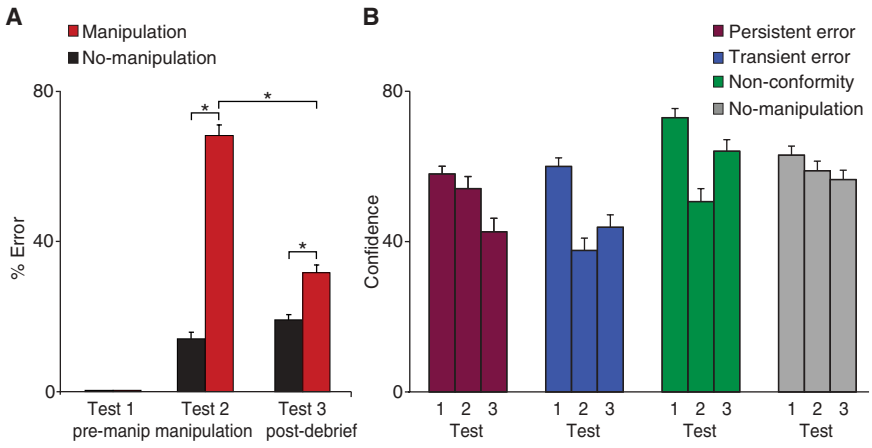

Fig. 2. Behavioral results. (**A**) Conformity level in the social manipulation condition was 68.3% versus 15.5% in the no-manipulation condition [t (19) = 16.9, $P < 10^{-7}$]. In test 3, participants reverted back to their original correct answer in 59.2% of the previously conformed-to events (transient errors) and on 40.8% maintained their erroneous answer (persistent memory error). The error rate was significantly different in test 3 between the manipulation and no-manipulation conditions [t (19) = 3.7, $P < 0.002$]. The questions included in the manipulation and no-manipulation trials were those for which participants gave correct answers in test 1 with medium-high confidence. (**B**) Confidence ratings over time for differential trial types (*$P < 0.002$).

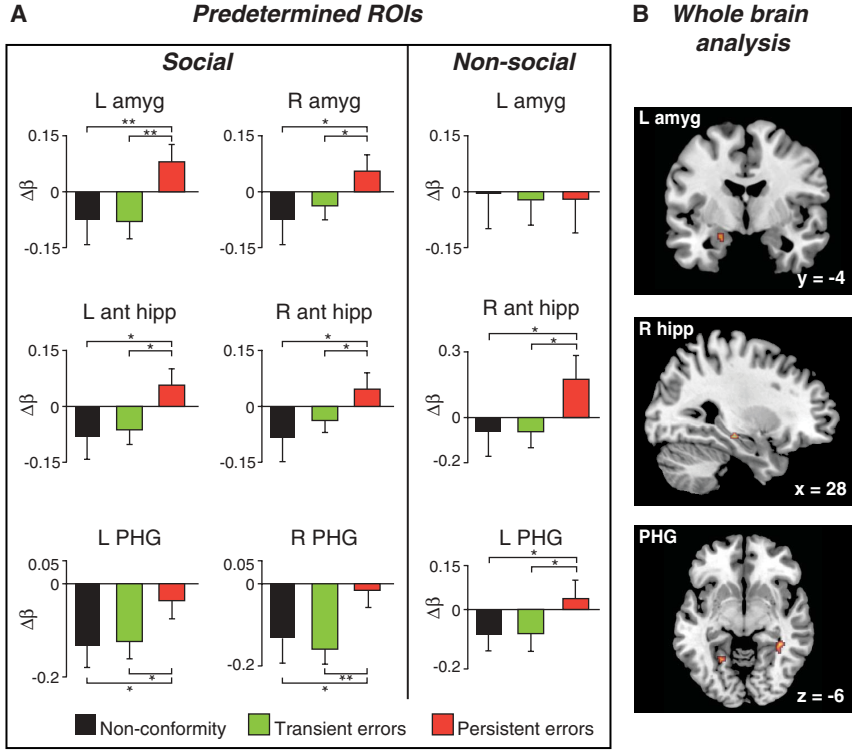

Fig. 3. MTL activation during manipulation predicts long-term socially induced memory errors. (**A**) BOLD signal in anatomically a priori defined MTL regions. L, left; R, right; In the social manipulation, enhanced activation was found during trials that subsequently resulted in persistent errors relative to all other conditions in the bilateral hippocampal complex and amygdala. In the nonsocial manipulation, this pattern was evident in the hippocampal complex but not in the bilateral amygdala. (**B**) Whole-brain exploratory analysis in the social manipulation ($P < 0.001$, $k > 10$) revealed greater activity in persistent error versus transient error conditions in the left amygdala, right hippocampus, right PHG, and left PHG bordering on the occipital lobe. All areas also survived small volume correction for multiple comparisons (familywise error < 0.05). The baseline in all figures is the no-manipulation condition (*$P < 0.05$ **$P < 0.005$).

Although at the time of social influence on memory overt behavior was indistinguishable, transient and persistent errors nevertheless in-

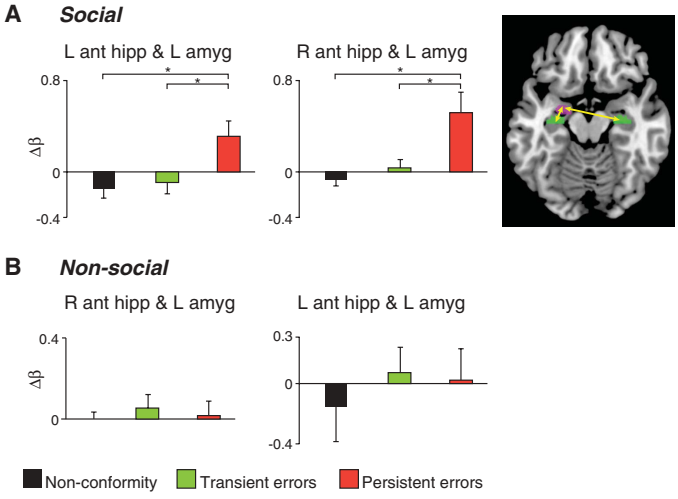

Fig. 4. Amygdala-hippocampal functional connectivity during manipulation predicts long-term socially induced memory errors. (**A**) Social manipulation. Functional connectivity between the left amygdala and bilateral anterior hippocampus was heightened in the persistent error condition relative to all other conditions. (**B**) Nonsocial manipulation. No condition-dependent difference in functional connectivity between the left amygdala and bilateral anterior hippocampus was found. The baseline in all figures is the no-manipulation condition (*$P < 0.05$). The inset depicts the anatomical ROIs used in the aforementioned analyses.

duced distinct brain signatures. Heightened activation in the hippocampal complex was seen when false information induced a long-lasting change in the participants' memories regardless of social context. The hippocampal complex activation we observed may represent a process of reconsolidation (*36*) or encoding of new stable representations [e.g., gist (*37*)]. In contrast, transient changes did not activate areas known to be crucial for memory processing. Our findings provide neurobiological evidence for the classic assertion that private conformity is accompanied by actual changes in beliefs, whereas public displays of conformity are not (*7*, *8*, *10, 38*).

Enhanced activation in the bilateral amygdala and heightened functional connectivity with the anterior hippocampus were a signature of long-term memory change induced by the social environment. This indicates that the incorporation of external social information into memory may involve the amygdala's intercedence, in accordance with its special position at the crossroads of social cognition and memory (*13*, *14*, *16*).

Multiple formal models have proposed trace attributes that might contribute to memory distortion in different false memory protocols (*37, 39–41*). These postulated attributes refer, for example, to potential heterogeneity in episodic content and the persistence of memory trace elements. Our laboratory analog to socially induced memory distortion was not intended to distinguish between specific models. However, further exploitation of our protocol, combined with cross-fertilization of behavioral and brain data, might contribute to the refinement of current models and better understanding of the biological and cognitive mechanisms of memory conformity.

Altering memory in response to group influence may produce untoward effects. For example, social influence such as false propaganda can deleteriously affect individuals' memory in political campaigns and commercial advertising (*1*, *2*, *6*) and impede justice by influencing eyewitness testimony (*2*, *4, 5*). However, memory conformity may also serve an adaptive purpose, because social learning is often more efficient and accurate than individual learning (*42*). For this reason, humans may be predisposed to trust the judgment of the group, even when it stands in opposition to their own original beliefs. Such influences and their long-term effects, the neurobiological basis of which we describe here, may contribute to the extraordinary levels of persistent conformity seen in authoritarian cults and societies.

References and Notes

1. M. L. Meade, H. L. Roediger 3rd, *Mem. Cognit.* **30**, 995 (2002).
2. E. F. Loftus, *Learn. Mem.* **12**, 361 (2005).
3. D. L. Schacter, *The Seven Sins of Memory: How the Mind Forgets and Remembers* (Houghton-Mifflin, New York, 2001).
4. D. B. Wright, A. Memon, E. M. Skagerberg, F. Gabbert, *Curr. Dir. Psychol. Sci.* **18**, 174 (2009).
5. J. S. Shaw 3rd, S. Garven, J. M. Wood, *Law Hum. Behav.* **21**, 503 (1997).
6. H. W. Perkins, J. W. Linkenbach, M. A. Lewis, C. Neighbors, *Addict. Behav.* **35**, 866 (2010).
7. E. Smith, D. Mackie, *Social Psychology* (Psychology Press, London, ed. 3, 2007).
8. V. Allen, in *Advances in Experimental and Social Psychology,* L. Berkowitz, Ed. (Academic Press, New York, 1965), pp. 133–170.
9. M. B. Reysen, *Memory* **13**, 87 (2005).
10. S. E. Asch, in *Groups, Leadership and Men*, H. Guetzkow, Ed. (Carnegie Press, Pittsburgh, PA, 1951), pp. 39–76.
11. L. Festinger, *A Theory of Cognitive Dissonance* (Peterson, Evanston, IL, 1957), pp. 99–100.
12. Y. Dudai, *Memory from A to Z. Keywords, Concepts and Beyond* (Oxford Univ. Press, Oxford, 2002).
13. R. Adolphs, *Nat. Rev. Neurosci.* **4**, 165 (2003).
14. E. A. Phelps, *Annu. Rev. Psychol.* **57**, 27 (2006).
15. F. Dolcos, K. S. LaBar, R. Cabeza, *Neuron* **42**, 855 (2004).
16. R. J. Dolan, *Science* **298**, 1191 (2002).
17. K. C. Bickart, C. I. Wright, R. J. Dautoff, B. C. Dickerson, L. F. Barrett, *Nat. Neurosci.* **14**, 163 (2010).
18. K. N. Ochsner, in *Social Neuroscience: People Thinking About People,* J. T. Cacioppo, Ed. (MIT Press, Cambridge, MA, 2005), pp. 245–268.
19. R. N. Cardinal, J. A. Parkinson, J. Hall, B. J. Everitt, *Neurosci. Biobehav. Rev.* **26**, 321 (2002).
20. L. R. Squire, *Neurobiol. Learn. Mem.* **82**, 171 (2004).
21. J. P. Aggleton, *The Amygdala: Second Edition. A Functional Analysis* (Oxford Univ. Press, Oxford, 2000).
22. H. Kluver, P. C. Bucy, *J. Psychol.* **5**, 33 (1938).
23. M. P. Richardson, B. A. Strange, R. J. Dolan, *Nat. Neurosci.* **7**, 278 (2004).
24. Y. Okado, C. E. L. Stark, *Learn. Mem.* **12**, 3 (2005).
25. L. Nadel, M. Moscovitch, *Curr. Opin. Neurobiol.* **7**, 217 (1997).
26. L. Stefanacci, D. G. Amaral, *J. Comp. Neurol.* **451**, 301 (2002).
27. K. J. Friston *et al.*, *Neuroimage* **6**, 218 (1997).
28. R. Cabeza, L. Nyberg, *J. Cogn. Neurosci.* **12**, 1 (2000).
29. E. K. Miller, J. D. Cohen, *Annu. Rev. Neurosci.* **24**, 167 (2001).
30. V. Klucharev, K. Hytönen, M. Rijpkema, A. Smidts, G. Fernández, *Neuron* **61**, 140 (2009).
31. D. M. Amodio, C. D. Frith, *Nat. Rev. Neurosci.* **7**, 268 (2006).
32. B. A. Kuhl, N. M. Dudukovic, I. Kahn, A. D. Wagner, *Nat. Neurosci.* **10**, 908 (2007).
33. B. J. Levy, M. C. Anderson, *Trends Cogn. Sci.* **6**, 299 (2002).
34. J. P. Mitchell, C. S. Dodson, D. L. Schacter, *J. Cogn. Neurosci.* **17**, 800 (2005).
35. F. C. Bartlett, *Remembering* (Cambridge Univ. Press, Cambridge, 1932).
36. Y. Dudai, *Annu. Rev. Psychol.* **55**, 51 (2004).
37. V. F. Reyna, C. J. Brainerd, *Learn. Individ. Differ.* **7**, 1 (1995).
38. M. Sherif, *The Psychology of Social Norms* (Harper Collins, New York, 1936).
39. H. L. Roediger 3rd, J. M. Watson, K. B. McDermott, D. A. Gallo, *Psychon. Bull. Rev.* **8**, 385 (2001).
40. J. Arndt, *Psychol. Learn.* **36**, 66 (2010).
41. M. L. Howe, *Psychol. Bull.* **134**, 768, discussion 773 (2008).
42. R. Boyd, P. J. Richardson, in *Social Learning: Psychological and Biological Perspectives,* R. R. Zentall, B. J. Galef, Eds. (Erlbaum, Hillsdale, NJ, 1988), pp. 29–48.

Acknowledgments: M.E. was supported by a Weizmann Institute–UK Grant. T.S. is supported by a British Academy Postdoctoral Fellowship. R.J.D is supported by a Wellcome Trust Program Grant. Y.D. is supported by the Nella and Leon Benoziyo Center for Neurological Diseases. We thank A. Ben-Yakov, J. G. Edelson, T. Fitzgerald, O. Furman, S. Fleming, D. Levi, M. Guitart-Masip, A. Mendelsohn, U. Nili, A. Pine, J. S. Winston, 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 unit in the Wellcome Trust Center for Neuroimaging.

Supporting Online Material

www.sciencemag.org/cgi/content/full/333/6038/108/DC1
Materials and Methods
Fig. S1
Table S1
References

31 January 2011; accepted 12 May 2011
10.1126/science.1203557

Supporting Online Material for

Following the Crowd: Brain Substrates of Long-Term Memory Conformity

Micah Edelson,* Tali Sharot, Raymond J. Dolan, Yadin Dudai

*To whom correspondence should be addressed. E-mail: micah.edelson@weizmann.ac.il

Published 1 July 2011, *Science* **333**, 108 (2011)
DOI: 10.1126/science.1203557

This PDF file includes:

Supporting Online Material

Methods

Task and stimuli

Participants: Thirty right handed subjects (12 females, average age 28.6 ± 0.77, (mean ± SEM)) participated in the study. One participant was omitted from further analysis because of suspected minor brain pathology and one due to head movements exceeding 4 mm. Only subjects who indicated no suspicion of the experimental manipulation when debriefed were included in the analysis (a final N = 20, 8 females, age 27.4 ± 1.0). All participants gave informed consent and were paid for their participation. The study was approved by the Institutional Review Board of the Sourasky Medical Center, Tel-Aviv.

***Stimuli*:** The stimuli consisted of a 40 minute eyewitness styled documentary following the activities of police deporting illegal immigrants. The film included scenes of forceful arrests of illegals and their families. The content had medium emotional valance as rated by participants (see results below).

Procedure (Fig. 1A)**:** The experiment was divided into four phases.

Encoding phase (day 0): The initial encoding of the movie was performed with groups of 5 unacquainted individuals. The participants introduced themselves 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* (day 3): Memory *Test 1* was a computerized 400 question, two-forced choice, memory questionnaire on the film's content, conducted three days after the encoding phase. After providing each answer the subjects rated how confident they were in their responses. The confidence ratings were provided 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. The average accuracy was 69.1 ± 1.2% for all answers and 80.2 ± 2.0% for answers with medium to high confidence scores.

Manipulation phase (day 7, Fig. 1A): Subjects performed a memory test while in an fMRI scanner. On each trial the participants were presented with a memory question related to the film. The questions were identical to those in memory *Test 1*; however, to minimize scanning time, only 320 questions were included (randomly selected). The question, two possible answers and pictures of the co-observers, who had seen the film together with the subject, were displayed for 2.5 seconds (mode of presentation adapted from *S1*). This was followed by a blank screen for a jittered 2 second interval (range: 1-8 seconds). The design allowed subjects to try and retrieve what they remembered before the false information was presented. Analysis for this time window is presented in supplementary results. Next, the manipulated co-observers answers were displayed on the screen for 2.5 seconds. The participants were not allowed to answer during this period to ensure that they gave due consideration to the new information presented. After the 2.5 second interval the color of the question font changed indicating to the subjects that they now could respond. They then provided a response and on 66% of the trials (randomly assigned) also provided a confidence rating. Participants were instructed that the answers of their co-observers could be used to assist their retrieval process but that they ultimately were required to answer according to their own recollection. The scan was divided into 3 sessions with a 15 minute break between sessions.

The co-observer answers were pseudo-randomly allocated into 3 different categories as follows: (Fig. 1B).

1. Manipulation condition. 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. 80 of these questions (randomly assigned) were included as manipulation questions in *Test 2*. The average confidence rating for all manipulation questions was 62.6 ± 2.3, lying between a medium (50) and high (75) confidence rating. In all manipulation questions the fake co-observer answers provided in memory *Test 2* were deliberately incorrect.

2. No-manipulation control condition. 25 different 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. 215 different questions were randomly chosen from all questions in memory *Test 1*. Since it is not credible that all co-observers answers would always be unanimously in disagreement with the participants' remembered view it was necessary to add additional questions in which the co-observers answers appeared in different patterns. Thus, the sole purpose of the *credibility condition* was to ensure that the manipulation questions were believable to the subjects. Credibility questions were not subsequently analyzed in the fMRI data. Pilot data showed that using less credibility questions led participant to suspect the manipulation. 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 correct answer the greater the number of correct answers given by co-observers.

Image acquisition and analysis

Image acquisition. Imaging was performed on a 3T scanner. 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 80º, 35 oblique slices without gap, 30º towards coronal plane from AC PC, 3 × 3 × 4 mm voxel size covering the whole cerebrum.

Image analysis. Statistical Parametric Mapping (SPM5; Wellcome Trust Centre for Neuroimaging, London, UK, http://www.fil.ion.ucl.ac.uk/spm) was used to analyze the fMRI data. After discarding the first 3 dummy volumes, images were realigned to the first volume, unwarped, normalized to a standard EPI template based on the Montreal Neurological Institute (MNI) reference brain, resampled to 2mm×2mm×2mm voxels, and spatially smoothed with an isotropic 8 mm full width at half maximum (FWHM) Gaussian kernel.

For each participant, a time series was created indicating the temporal position of the different trial types. Data for individual trial types were convolved with the canonical hemodynamic response using a random effect general linear model (GLM). For the GLM, 11 regressors were constructed. Of these, 5 regressors were created for the critical time period when co-observer answers' were presented (results of this time window are included in the main text): a. *Persistent errors*: Manipulation trials for which participants initially answered correctly (*Test 1*) but gave incorrect answers in both *Test 2* and *Test 3*. b. *Transient errors*: Manipulation trials for which participants initially answered correctly (*Test 1*) and gave incorrect answers in *Test 2* but not in

Test 3. c. *Non-conformity*: Manipulation trials for which participants gave a correct answer in both *Test 1* and *Test 2*. d. *No-manipulation condition* questions. e. *Credibility condition* questions. These regressors were modeled as a boxcar from the time the co-observers' answers were presented until the participant responded. We allowed this boxcar to reach a maximum of 6.5 seconds. The 6.5 second maximum was included in order to focus our analysis on the initial time frame of social influence on the participants' decision (in 18.6 ± 2.3% of the trials participants' reaction was given after this time frame). Five additional regressors were created for the period of the question presentation. These 5 regressors again corresponded to the 5 experimental conditions (*persistent errors*, *transient errors*, *non-conformity*, *no-manipulation* and *credibility*). These regressors were modeled as a 2.5 sec boxcar. An additional regressor was created for the time window of the confidence rating phase and was modeled as a boxcar from the time of presentation of the VAS confidence scale until response. Note that differences in confidence ratings and reaction times were controlled for by adding a vector including each subject deferential values as covariates in the second level analysis for all contrasts.

Region of interest analysis. For all region of interest (ROI) analysis we extracted the mean parameter estimates averaged across the whole ROI for each experimental condition separately and entered them into a repeated measures ANOVA analysis with experimental condition (*persistent errors, transient errors, non-conformity* and *no-manipulation*) as a factor. When significant, this was followed by t-tests. Two types of ROI were used in our analysis:

a. *A-priori anatomical ROIs:* The a-priori anatomical ROIs were defined based on known anatomical landmarks according to the Talairach Daemon Atlas (*S2*) using the SPM WFU PickAtlas tool (*S3*). Anatomical ROIs were defined for the bilateral amygdala, bilateral para-hippocampus and bilateral anterior and posterior hippocampus. The hippocampus subdivision was defined according to previous literature (*S4*).

b. *Functional ROIs:* Functional ROIs were defined in the *social manipulation* experiment. The ROIs were regions that showed increased activation when additional information was present (*manipulation condition* > *no-manipulation condition*, $p < 0.00005$ $k > 50$). For the *non-social manipulation* experiment we identified voxels where activity was grater in the *non-social manipulation* vs. the *no-manipulation conditions* within the functional ROIs identified above (small volume correction (SVC), FWE < 0.05).

Functional connectivity analysis. A whole-brain psychophysiological interaction (PPI) analysis was conducted to identify if target brain regions showed a significant difference in functional coupling with the left amygdala in the different conditions of interest (i.e. *persistent errors*, *transient errors, non-conformity* and *no manipulation condition*). Our target regions were the anatomically defined hippocampal complex ROIs. The regressors in the PPI analysis included: 1. The activation time course of the volume of interest (i.e. physiological variable; the BOLD signal). 2. A regressor representing the psychological variable of interest (i.e. the different experimental conditions). 3. A regressor representing the cross product of the previous two (the psychophysiological interaction term, PPI). The first 2 regressors were added as covariates to the model whilst the last regressor was the regressor of interest. For each subject, we averaged the parameter estimates of the PPI regressor across the whole target ROI for each condition of interest separately and conducted a repeated measures ANOVA analysis with experimental condition (*persistent errors, transient errors, non-conformity* and *no-manipulation*) as a factor (for previous literature on method see *S5*-*S6*). When significant, this was followed by t-tests. Fig. 4A and Fig. 4B present functional connectivity results for the social and non-social experiment respectively.

Supplementary Results

Behavior

Confidence ratings. Confidence ratings in the *persistent* and *transient error*s did not differ before or after the manipulation stage (Fig. 2B). During the manipulation stage confidence ratings in *transient errors* dropped significantly, and were lower than for *persistent errors* (t (19) = 6.8, $p < 10^{-5}$). This may indicate that in the former case, while publicly conforming, participants still considered their response to be incorrect, whilst in the latter case they accepted the new information without conflict (*S7*). When social influence was removed, confidence levels for both *persistent* and *transient errors* reconverted to medium confidence levels. In the *non-conformity* trials, confidence dropped significantly as the participants maintained their opinion despite social pressure (*Test 2*), but increased again when social influence was removed (t (19) = 8.1 and t (19) = 4.7 respectively, $p < 0.0002$). Note that differences in confidence were controlled for in the fMRI analyses using a covariate.

Reaction times (RT). A repeated measures ANOVA, using experimental condition as a factor, revealed a significant difference in RTs (F (3, 57) = 6.1 p < 0.001). Shorter RT's were found in the *no-manipulation condition* compared to the 3 other conditions (i.e. *persistent errors*, *transient errors* and *non-conformity*), ($p < 0.002$). Longer RT's were found for *non-conformity* compared to *persistent error*s (t (19) = 2.5, $p < 0.02$). This pattern of reaction time may indicate an increasing level of conflict (*S8*). As aforementioned, these differences were controlled for by adding the deferential RT as a covariate in the second level analysis.

Debriefing results. Eight subjects were excluded from the analysis because they indicated suspicion that the co-observers answers presented in memory *Test 2* were fabricated. Subjects were excluded even if they indicated that their suspicion was weak and did not affect their behavior. We used this conservative inclusion threshold in order to avoid confounds related to uncovering the manipulation. Analysis of the excluded subjects revealed that they had significantly less conforming behavior compared to the included subjects (49.0 ± 4.5%, for excluded participants vs. 68.3 ± 2.9% for included participants t (26) = 3.2, $p < 0.005$). Memory performance in *Test 1* did not differ between excluded and included participants (t (26) = 1.7, $p > 0.1$).

All 20 participants included in the final analysis indicated that on memory *Test 3* they understood that the co-observer information viewed previously (on memory *Test 2*) was irrelevant. They indicated that, as instructed, they attempted on every trial to answer only from their own memory of the original movie. Consistent with previous studies (*S9*), debriefing indicated that participants regarded the *persistent memory errors* as vivid personal experiences. None of the participants were consciously aware of incorporating information provided by the co-observers into their own recollection. However, all 20 participants were aware that they sometimes reverted back to their original answer in the final test. Seventeen of the 20 participants indicated that this latter circumstance occurred only when they had "publicly" conformed.

Functional imaging

Brain activation during manipulation condition* vs. *no-manipulation condition.

Five brain regions showed enhanced activation during *manipulation* relative to *no-manipulation conditions* (Fig. *S1A* and Table *S1*, $p < 0.00005$, k > 50). These included four frontal regions; bilateral inferior frontal gyrus (IFG; BA 47; 32,22,-14; -32,16,-20), dorsal ACC (BA 32; 10,32,34), dorsal medial pre-frontal cortex (dmPFC, BA 8; 6,24,50) and an additional region in

the occipital cortex (BA 17; -10,-94,-6). Averaging activity in these regions revealed larger activity in all frontal regions during trials in which the participant did not conform as opposed to trials where they conformed (Fig *S1B*). There was no difference between trials that resulted in long lasting memory change and those resulting in only *transient errors* (with one exception in dmPFC). Conjoint activation in these frontal areas has been associated with conflict detection and cognitive control (*S10-S12*), such as when confronted by competition from irrelevant memories (*S13-S14*). Thus in contrast to the MTL, activity here may reflect the explicit decision of the participant not to conform or rising conflict levels, rather than long term memory modulation (this interpretation was supported by longer reaction time in the *non-conformity* condition). The occipital cortex region was found, presumably, because more complex visual stimulation was present during the *manipulation condition* (co-observers text answers were displayed) relative to the *no-manipulation condition* (only letter 'X' was displayed).

Brain activation during question presentation. The experimental protocol was designed to allow the participant to recall the information on their own before being exposed to social influence. To this end the question and possible answers were displayed for 2.5 second without the co-observers answers. Whole brain analysis during this time window ($p < 0.001$, $k = 10$) did not reveal any significant differences between the *persistent errors, transient errors* and the *non conformity* conditions.

Control Study I: Emotional arousal

Activation in the amygdala associated with emotional arousal has been repeatedly demonstrated in the literature (*S15-S17*). Ratings of emotional arousal have been previously correlated with independent measures of physiological arousal such as skin conductance response and amygdala activation (*S18*). To examine whether the elevated activation in the amygdala during trials that resulted in *persistent memory errors* were related to heightened emotional arousal we conducted an additional behavioral study collecting ratings of emotional arousal on a trial by trial basis from our participants on all three memory tests.

Design: 10 participants (6 females, average age 26.2 ± 1.3) were tested in the same behavioral protocol used in the main experiment with the following additions; in memory *Test 1* the participants were asked to retrospectively rate the emotional arousal they felt while watching the part of the movie associated with the question. In memory *Test 2* and memory *Test 3* the participants were asked to rate the emotional arousal they felt at the present moment. The participants rated using a VAS scale ranging from 0 (no emotional arousal) to 100 (very high emotional arousal), with 25 indicating low emotional arousal, 50 medium emotional arousal and 75 high emotional arousal.

Results: We preformed a repeated measures ANOVA analysis with experimental condition (*persistent errors, transient errors, non-conformity* and *no-manipulation*) as a factor. The average emotional rating in memory *Test 1* was 51.2 ± 2.3 indicating that the movie content was perceived as emotional on a medium level. We found no significant differences in emotional arousal ratings between the different conditions in memory *Test 2* or memory *Test 3*. In memory *Test 1* there was a significant effect (F (3,27) = 4.7, $p < 0.02$) which was driven by higher ratings for questions that will result in *non-conformity* relative to *transient errors* (t (9) = 3.2, $p < 0.05$), *persistent errors* (t (9) = 2.1, $p < 0.06$) and *no-manipulation* (t (9) = 2.8, $p < 0.05$). These results are consistent with previous literature demonstrating that highly emotional material is less likely to undergo

conformity (*S19*) and more likely to be accurately remembered (*S20*). Importantly, however, no difference was found between the *persistent* and *transient error* conditions.

Control Study II: Non-social manipulation

The question arises as to whether our findings are driven by a unique social context or rather demonstrate a more generalized reaction to misinformation (S*20-S22*). To this end we performed a control experiment using a non-social medium to convey misinformation, a technique commonly used for this purpose (*S1,S22-S23*).

Design: 20 participants (9 females, 28.1 ± 1.1) underwent a similar protocol to the one in our main experiment. However, in memory *Test 2*, instead of receiving answers from co-observers, subjects received the same information but were told that it originated from 4 different computer algorithms. The participants were told that the different algorithms had been tested and proven to provide an accuracy level equal to, and sometimes slightly higher than that of humans. Three participants were excluded from the analysis because they indicated suspicion of the manipulation, three participants were excluded due to technical problems and one participant was excluded due to claustrophobia in the scanner setting (final N of 13).

Behavioral Results: The conformity levels when answers were given by computers (*non-social manipulation condition*, 45.3 ± 4.7%), was significantly lower than in the *social manipulation condition* described in the main text (68.3 ± 2.9%), but significantly higher than when *no-manipulation* at all was presented (15.0 ± 2.4%), (t (38) = 4.2 and t (19) = -5.7 respectively, $p <$ 0.0002). When influence was removed (*Test 3*) in this non-social control, participants reverted back to their original, correct answer in 61.1 ± 2.6% of the previously conformed trials (*transient errors*), but maintained erroneous answers in 38.9% (*persistent errors*). There was no interaction between manipulation (*social/non-social*) and type of error (*persistent/transient*), ($p > 0.6$), suggesting that social manipulation increased both types of errors equally.

Functional imaging results

a. Persistent vs. *transient errors.* Greater activation during trials that resulted in *persistent* relative to *transient errors* (whole brain analysis, $p = 0.001$ $k = 10$) was found in 2 MTL regions; the left PHG and the right hippocampus (-24,-48,-12, $t = 4.6$ and 34,-20,-19, $t = 4.2$ respectively). Significant activation was also found in the caudate nucleus (26,20,32, $t = 6.9$) and the occipital cortex (30,-98,12, $t = 6.9$). No activation was found in the amygdala even when applying SVC. The reverse contrast (*transient errors > persistent errors*) did not reveal any significant result.

b. Non-social manipulation vs. no-manipulation condition. We examined whether the 5 regions identified in the main experiment as differentiating between *manipulation* and *no-manipulation* trials (Fig. S1A, Table S1) show the same pattern of activation in this control experiment. We contrasted *non-social manipulation* trials with *no-manipulation* trials in these ROIs. In all regions but one, enhanced activation was found during the *non-social manipulation condition* ($p < 0.05$, SVC, FWE). The regions were; right IFG (BA 45; 54,24,6), dMPFC (BA 8; 4,36,48), ACC (BA 32/9; peak voxel in 4,36,38) and the occipital cortex (BA 17; -2,-92,-4). Averaging activity in these regions **(**Fig. S1B) revealed that frontal areas showed heightened activation for the *non-conformity* trials relative to all other trials. No such differences were found in the occipital region. Thus activity in these regions demonstrated a similar pattern in the *social* and *non-social manipulations* consistent with a proposed role in non-specific conflict monitoring and decision making.

Supplementary References

S1. G. S. Berns, J. Chappelow, C. F. Zink, G. Pagnoni, M. E. Martin-Skurski, J. Richards, Neurobiological correlates of social conformity and independence during mental rotation. *Biol. Psychiatry* **58**, 245-253 (2005).

S2. J. L. Lancaster, L. H. Rainey, J. L. Summerlin, C. S. Freitas, P. T. Fox, A. C. Evans, A. W. Toga, J. C. Mazziotta, 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 (1997).

S3. J. A. Maldjian, P. J. Laurienti, R. A. Kraft, J. H. Burdette, An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* **19**, 1233-1239 (2003).

S4. A. Gilboa, G. Winocur, C. L. Grady, S. J. Hevenor, M. Moscovicth, Remembering our past: functional neuroanatomy of recollection of recent and very remote personal events *Cereb. Cortex* **14**, 1214-1225 (2004).

S5. K. J. Friston, C. Buechel, G. R. Fink, J. Morris, E. Rolls, R. J. Dolan, Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* **6,** 218-229 (1997).

S6. T. Sharot, T. Shiner, R. J. Dolan, experience and choice shape expected aversive outcomes. *J Neurosci.* **30**, 9209-9215 (2010).

S7. C. A. E. Luus, G. L. Wells, The malleability of eyewitness confidence: co-witness and perseverance effects. *J. Appl. Psychol.* **79**, 714-723 (1994).

S8. B. A. Eriksen, C. W. Eriksen, Effects of noise letters upon the identification of a target letter in a non-search task. *Percept. Psychophy* **16**, 143-149 (1974).

S9. E. F. Loftus, K. Donders, H. G. Hoffman, J. W. Schooler, Creating new memories that are quickly accessed and confidently held. *Memory Cognition* **17**, 607-617 (1989).

S10. R. Cabeza R, L. Nyberg, Image cognition II. An empirical review of 275 PET and fMRI studies. *J. Cogn. Neurosci.* **12**, 1-47 (2000).

S11. E. K. Miller, J. D. Cohen, An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci*. **24**, 167-202 (2001).

S12. B. J. Levy, M. C. Anderson, Inhibitory processes and the control of memory retrieval. *Trends Cogn. Sci*. **6**, 299-305 (2002).

S13. B. A. Kuhl, N. M. Dudukovic, I. Kahn, A. D. Wagner, Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nat. Neurosci*. **10**, 908-914 (2007).

S14. J. P. Mitchell, C. S. Dodson, D. L Schacter, fMRI evidence for the role of recollection in suppressing misattribution errors: the illusory truth effect. *J. Cogn. Neurosci.* **17**, 800-810 (2005).

S15. E. A. Phelps, Human emotion and memory: interactions of the amygdala and hippocampal complex. *Curr. Opin. Neurobiol.* **14**, 198-202 (2004).

S16. R. Adolphs, Is the human amygdala specialized for processing social information? *Ann NY Acad Sci* ***985***, 326-340 (2003).

S17. K. N. Ochsner, in *Social Neuroscience: People thinking about people*, J. T. Cacioppo, Ed. (MIT Press, Cambridge, 2005), pp. 245-268.

S18. S. B. Hamann, T. D. Ely, S. T. Grafton, C. D. Kilts, Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nat. Neurosci.* **2,** 289-293 (1999).

S19. C. Brown, A. Schaefer, Effects of conformity on recognition judgments for emotional stimuli. *Acta Psychol* **133** ,38-44 (2010).

S20. E. A. Phelps, Emotion and cognition: insights from studies of the human amygdala. *Annu. Rev. Psychol* ***57***, 27-53 (2006).

S21. J. P. Mitchell, C. N. Macrae, M. R. Banaji, Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *J. Neurosci* **24** 4912-917 (2004).

S22. R. Adolphs, The social brain: neural basis of social knowledge. *Annu. Rev. Psychol.* **60**, 693-716 (2009).

S23. V. Klucharev, K. Hytönen, M. Rijpkema, A. Smidts, G. Fernández, Reinforcement learning signal predicts social conformity. *Neuron* **61**, 140-151 (2009).

Fig. S1.

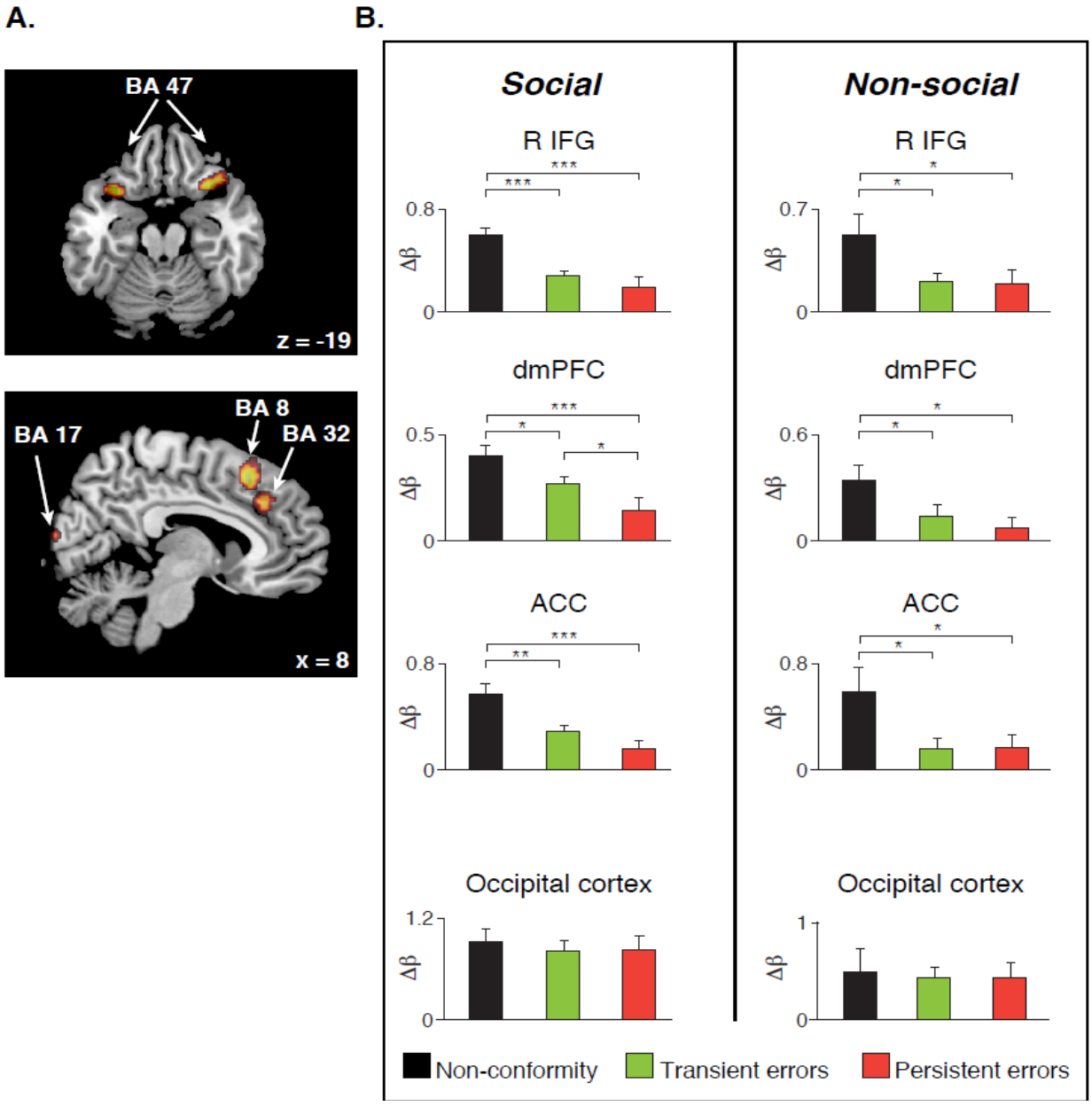

Fig. S1. **(A)** Regions identified by contrasting activation during the *social manipulation condition* relative to *no-manipulation condition* ($p < 0.00005$ uncorrected, $k > 50$, all areas also survived FWE $p < 0.05$ whole brain corrected); bilateral BA 47, BA 32, BA 8 and BA 17. **(B)** Of these regions, activity in frontal areas was greater in the *non-conformity* condition than either conformity conditions in both the *social* and *non-social manipulations*. The occipital cortex showed heightened activation for all conditions in which text answers were displayed regardless of the social context. The baseline in all figures is the *no-manipulation condition*. (* $p < 0.05$ ** $p < 0.005$ *** $p < 0.0005$)

Table S1 *Whole brain analysis in social experiment* (*social manipulation vs. no-manipulation conditions).*

Region	MNI			*t*	*p* (FWE corrected for whole brain)
	X	Y	Z		
Bi-lateral inferior frontal gyrus (peak at BA 47; extending into the ventrolateral prefrontal cortex)	32, -32	22 16	-14 20	9.0 7.8	0.004 0.019
Dorsal ACC (peak at BA 32, extending into the rostral ACC at a slightly lower threshold)	10	32	34	7.3	0.04
dmPFC, (peak at BA 8; extending into BA 6 and 32)	6	24	50	8.1	0.014
Occipital cortex (peak at BA 17)	-10	-94	-6	7.3	0.042