Covert neurofeedback without awareness shapes cortical network spontaneous connectivity

Michal Ramot[a,1], Shany Grossman[a], Doron Friedman[b], and Rafael Malach[a]

[a]Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel; and [b]Sammy Ofer School of Communication, Interdisciplinary Center, Herzlia 4610101, Israel

Edited by Marcus E. Raichle, Washington University in St. Louis, MO, and approved March 10, 2016 (received for review August 24, 2015)

Recent advances in blood oxygen level-dependent–functional MRI (BOLD-fMRI)-based neurofeedback reveal that participants can modulate neuronal properties. However, it is unknown whether such training effects can be introduced in the absence of participants' awareness that they are being trained. Here, we show unconscious neurofeedback training, which consequently produced changes in functional connectivity, introduced in participants who received positive and negative rewards that were covertly coupled to activity in two category-selective visual cortex regions. The results indicate that brain networks can be modified even in the complete absence of intention and awareness of the learning situation, raising intriguing possibilities for clinical interventions.

neurofeedback | training | reward | spontaneous activity | functional connectivity

There has been a growing interest in the field of neuroscience in the use of neurofeedback (NF) as a tool to both study and treat various clinical conditions. The uses of NF are diverse, ranging across a variety of motor and sensory tasks (1–4), investigation of cortical plasticity and attention (5–9), to treatment of chronic pain, depression, and mood control (10–13).

Recent advances in functional MRI (fMRI) techniques and hardware have made real-time fMRI (rtfMRI) a viable method for NF (14). This enables more anatomically specific training compared with methods such as EEG. This enhanced localization additionally allows to provide feedback to differential activation patterns (6, 15, 16), beyond simple up/down-regulation of a specific region/frequency.

Another advance in the field of NF is the finding by several recent studies that participants are able to learn to successfully perform the NF paradigm, even without being given an explicit strategy (8, 16). This form of implicit learning is intriguing, both because there have been reports indicating certain advantages to implicit over explicit learning (17, 18), but mostly because this opens up previously unidentified pathways for therapeutic intervention, for cases for which there are no specific explicit strategies available (for instance, control over complex networks, such as in epilepsy, or over brain regions whose function is not fully elucidated).

However, an important common factor in all previous NF studies was the fact that participants were aware that they were being trained, and received specific goals for this training. A fundamental question that therefore remains unanswered is whether targeted brain networks can still be modulated even in the complete absence of participants' awareness that a training process is taking place. Theories of closed-loop learning provide evidence that such implicit learning through reward cues is possible (19, 20). This is an important issue, because it may open the way for NF training even in severe clinical cases such as minimally conscious or vegetative state, where such awareness is absent.

In the present study, we examined this question in fMRI experiments in which participants were informed that they were engaged in a task aimed at mapping reward networks. Unbeknownst to them, these rewards were coupled with fMRI activations in specific cortical networks. Participants received auditory feedback associated with positive and negative rewards, based on blood oxygen level-dependent (BOLD)–fMRI activity from two well-researched visual regions of interest (ROIs), the fusiform face area (FFA) and the parahippocampal place area (PPA) (21–23). However, participants were not informed of this procedure and believed, as revealed also by postscan interviews and questionnaires, that the reward was given at random.

We have examined whether participants could learn implicitly to appropriately modulate their spontaneous cortical activity to increase reward. Previous work in our group (24) and others (25–27) has demonstrated that training effects, albeit with explicit participants' awareness of the training procedure, may leave a trace in the spontaneous patterns. Our question was whether such a trace could be similarly found following our covert training, with the crucial difference being that here participants had no explicit knowledge of the NF task, or even that it was possible to influence the reward.

Our results show that 10 of 16 participants (62.5%) were indeed able to modulate their brain activity to enhance the positive rewards. Importantly, participants were completely unaware that they were so doing. We further show that this ability was associated with changes in connectivity that were apparent in the posttraining rest sessions, indicating that the network changes resulting from the training carried over beyond the training period itself.

Results

Implicit NF. A total of 18 participants was enrolled in this study. Two were removed from the experiment after the first day of scanning, due to excessive movement (*Methods*). The remaining 16 participants were scanned on 5 separate days, within a 1-wk time frame. Before the start of the experiment, participants were

Significance

Real-time functional MRI allows the use of well-localized, complex network activity patterns to drive neurofeedback, rather than a simple up/down regulation of a specific cortical region. We based our feedback on differential levels of activity in two high-order visual areas but misled participants to believe the feedback was random. Even without being given an explicit strategy, or having any awareness or intention of learning, our results show changes in resting-state connectivity, which are correlated with the ability to implicitly modulate interactions between neural networks to positively impact feedback. This opens up numerous possibilities for research, as well as for potential clinical intervention, even in states of altered consciousness.

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

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: michal.ramot@nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516857113/-/DCSupplemental.

randomly assigned to either the "FFA-positive" or the "PPA-positive" group (hence FFA/PPA group). FFA and PPA were identified from an independent localizer, which was collected before the beginning of the experiment in a way that would avoid any association between the localizer and the main experiment (*Methods*). Sessions were identical across days and were composed of the following scans: (*i*) a 9-min rest scan, for which participants were instructed to simply remain still with their eyes closed; (*ii*) five consecutive NF scans, each 10 min long; (*iii*) a final 9-min rest scan, for which again participants were instructed to rest with their eyes closed (Fig. 1).

During the NF scans, for each repetition time (TR), participants received either positive, negative, or no auditory feedback sounds. Feedback was determined by an algorithm that compared activity levels (relative to baseline) in the FFA with those in the PPA, separately for each TR. For the FFA group, positive feedback was given if the FFA was activated above the PPA and over a certain threshold, whereas negative feedback was given if the PPA was more active than the FFA (using the same threshold). If the ratio of FFA/PPA activation was below threshold, no feedback was given. This was the same for the PPA group, with the roles of FFA and PPA reversed (*Methods*). In order for the participants to remain engaged, and yet still have ample room for improvement, the threshold was set so that either positive or negative feedback would be received in roughly one-third of the TRs, which with our scanning parameters translated to ∼100 feedback events for each scan. On average across all scans and subjects, there were 5.9 TRs between positive tones, and 6.2 TRs between negative tones. Overall, 36.4% of positive tones and 35.5% of negative tones were consecutive (i.e., following another positive/negative tone). Participants were instructed to lie in the scanner with their eyes closed, and press one button on the response box for each positive-feedback sound, and another button for negative-feedback sounds.

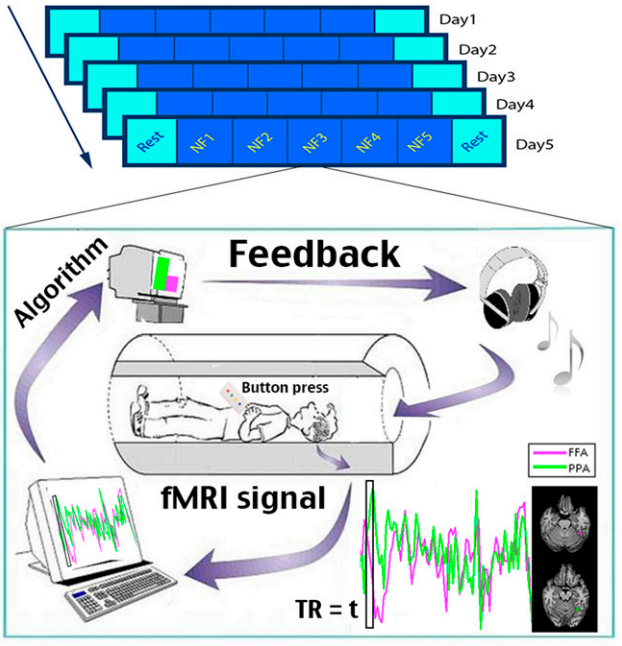

Fig. 1. Experimental paradigm. Participants were scanned for 5 d, with each day composed of an initial rest scan, five NF scans, and a final rest scan. During the NF, activity levels in our two chosen ROIs were contrasted in each TR, and participants were provided with auditory feedback based on the output of an algorithm that calculated the differential activation for these two ROIs. Feedback could be either positive, negative, or none. Participants were instructed to respond with a button press whenever they heard either a positive or negative beep.

Importantly, participants were not informed that the reward depended on their own brain activity. Rather, participants were told that the experiment was aimed at mapping positive and negative reward responses and that they would receive monetary compensation for each positive sound and would lose a similar amount for each negative sound. The sounds were chosen to be inherently associated with good/bad connotations (similar to computer game win or lose sounds). This created an implicit incentive for participants to wish for positive feedback while avoiding negative feedback, without knowing that the sounds reflected the activity levels in their visual cortex. The button-press task was designed to maintain alertness and attentiveness to the feedback. Participants were told that they must press the correct button after each positive/negative sound, or else they would lose the monetary reward (in the case of the positive feedback) or would be further monetarily penalized (in the case of the negative feedback).

In postexperiment questionnaires, participants were found to have no knowledge of the origin of the feedback sounds. Eleven of 16 participants thought the feedback sounds were random, 4 of 16 thought they might be based on their button-press response times or their mood, and only 1 participant suspected they might be some kind of NF based on other cortical activity. There was no correlation between success rates and participants' beliefs as to the nature of the feedback. When told that the sounds were indeed NF based on activity in two cortical regions, none of the participants had any strong notion of what the feedback might be correlated with. When presented with a five-alternative forced choice (*Methods*), participants were at chance in correctly identifying source of the feedback, and explicitly said that they were guessing.

Mapping the Algorithm's Effect onto the Brain. To examine the differential BOLD response to the positive vs. negative reward feedback sounds, we constructed a protocol based on the timing of these feedback sounds. Fig. 2 shows the random effects group analysis of the positive > negative contrast of the general linear model (GLM) built on this protocol, when adding the typical hemodynamic response of 6 s. This represents the differential response to the feedback sounds themselves, as opposed to the events that triggered the feedback. This figure shows that auditory cortex was more strongly activated for positive vs. negative sounds, which may be related to differential saliency of the positive vs. negative audio cues. More interestingly, however, is the activation found in the reward network-associated caudate body and lentiform nucleus, which may be related to the positive reward aspects of the sounds.

Because of the slow nature of the BOLD response, the feedback was actually given on neural activity that had taken place ∼6 s earlier (*Discussion*). To visualize the activity that triggered the feedback, we used the protocol based on the timing of the feedback sounds, but without adding the typical hemodynamic response function. Fig. S1 shows the results of the random-effects group analysis of the positive > negative contrast of the resulting GLM. It is apparent that, apart from lower activation in the "bad" ROI, there was widespread activation in networks associated with the "good" ROI, as well as increased activation in the thalamus, cerebellum, and posterior cingulate cortex. In other words, the activation in the brain at the time the feedback was produced was of a widespread nature, rather than being limited only to the targeted ROIs.

Modulation of Network Activity Following NF. To assess whether the covert NF training actually elicited significant changes in the relative FFA/PPA activation levels, we examined the algorithm's output (which for each TR, could be positive, negative, or neutral)

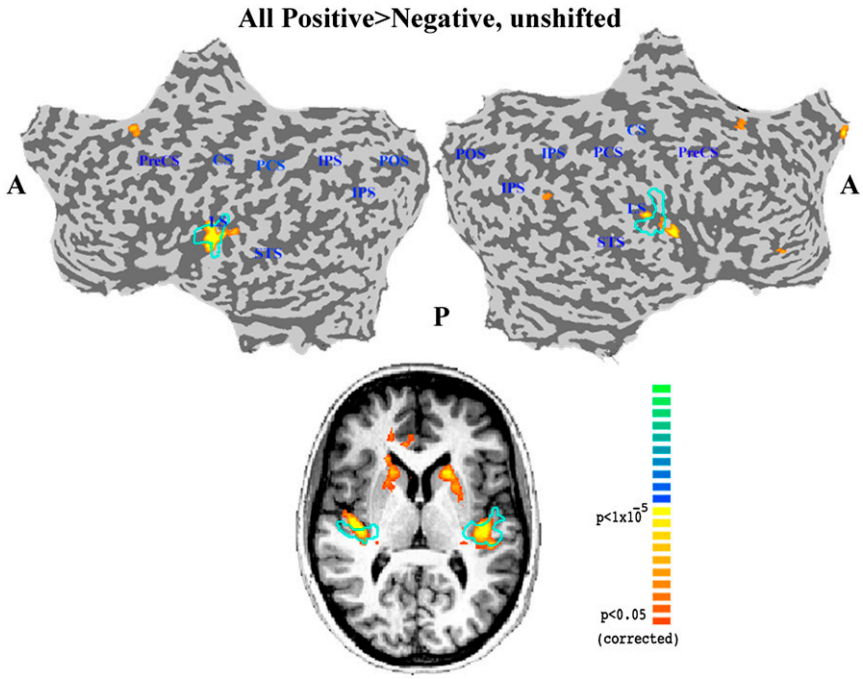

Fig. 2. Mapping the reward. The GLM was built on the protocol based on the feedback events, so that it corresponds to the effect of the feedback sounds. The maps show the positive > negative contrasts, for all subjects. The activation in auditory cortex (marked by cyan outline) may be related to the reward, or perhaps to the increased saliency of the positive feedback. Note the activation in the caudate nucleus and the putamen, both known to be involved in reward. A, anterior; CS, central sulcus; IPS, inferior parietal sulcus; LS, lateral sulcus; P, posterior; PCS, postcentral sulcus; POS, precuneus; PreCS, precentral sulcus; STS, superior temporal sulcus.

both during the NF sessions and the rest sessions before and after the training. During NF, positive/negative algorithm output triggered positive/negative-feedback sounds, whereas a neutral output triggered no feedback. For rest sessions, the same algorithm was used in the same way as with NF; however, its output was not set to trigger any feedback. In this manner, we obtained a record of the algorithm's output for the rest sessions, calculated in exactly the same way as for the NF, with the only difference being that the participants did not receive feedback on this output. For each session, we examined the number of positive algorithm outputs as a percentage of overall total positive-plus-negative output (positive/positive + negative). Successful participants were defined as those with over 50% positive outputs accumulated across all of the NF sessions. Ten such successful participants were identified from our data, six belonging to the FFA group, and four to the PPA group. There was no significant difference between the total number of feedback events (positive + negative) between the successful and unsuccessful participants. Although the number of successful participants is not in itself significant, there were a number of critical differences between the successful and unsuccessful participants.

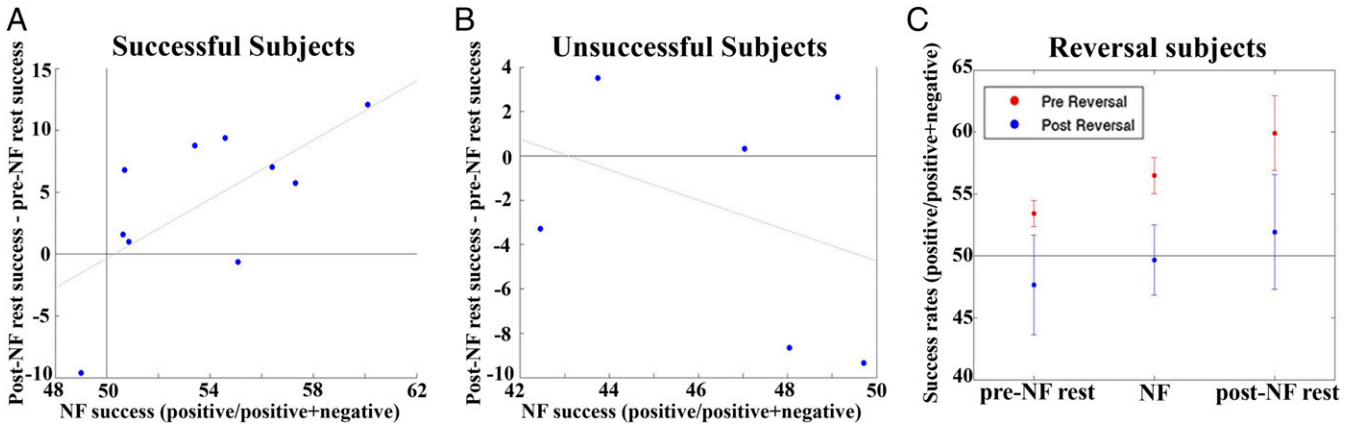

Fig. 3. Within-session learning effects. Each dot represents the difference in success rate between the post-NF rest scan and the pre-NF rest scan according to the algorithm's output (positive/positive + negative) for the average of the first 3 scanning days, vs. the averaged success rate of the five NF scans for those days, for one participant. Data for the successful participants are shown in *A*, whereas data for the unsuccessful participants is shown in *B*. (*C*) Results of the reversal analysis. The mean and SEM (calculated between subjects) of the success rate (positive/positive + negative) for each condition (pre-NF rest, NF, post-NF rest) is shown for the four successful reversal subjects. The average of the first 3 prereversal days shown in red, and the average of the last 2 postreversal days is shown in blue.

In each day, successful participants showed improved performance for the training sessions compared with the pretraining rest. Importantly, although there was no significant change in performance between days, either between the NF sessions or the pretraining rest sessions, there was a consistent and significant improvement from the pre-NF to the post-NF rest for each day, for the successful subjects only. It should be noted that, as described above, the successful participants were chosen based on their performance during the NF sessions only, which should not in itself predict this change between the rest sessions. Fig. 3 *A* and *B* shows this difference in the algorithm's output for the post-NF rest minus pre-NF rest, vs. the algorithm's output for the NF session each day, for the successful and unsuccessful participants, respectively. For each subject, we used the average of the first 3 prereversal days (see below for explanation of reversal condition). The successful subjects showed a positive and significant correlation between the change in rest performance and the NF success for each day for these days ($r = 0.67$, $P < 0.017$, calculated by permutation test). Because it is impossible to reliably calculate significance for the unsuccessful subjects using only six data points, we repeated the analysis using all of the data points, although this has the problem of combining both within-subject and between-subject correlations. We again found a significant correlation for the successful subjects ($r = 0.41$, $P = 0.027$, permutation test), but no such trend for the unsuccessful subjects ($r = -0.15$, $P = 0.68$, permutation test).

To assess the statistical significance of this increase from pretraining to NF, and from pretraining rest to posttraining rest, we performed a permutation test, shuffling daily session labels 1,000 times. For the successful participants, across both reversal conditions and all days, these changes were significant ($P < 0.012$ for NF vs. pre-NF rest, $P < 0.015$ for post-NF rest vs. pre-NF rest). A similar analysis on the unsuccessful participants revealed no such significant difference.

Of the 16 participants, 6 participants were assigned to a reversal condition, meaning that on the fourth of five training days they were switched to the alternative (FFA/PPA) group. This reversal condition was aimed at neutralizing any potential baseline bias effects that might occur in each individual participant, randomly predisposing that participant to more positive/negative outcome regardless of the training. If these results were a product of such a simple baseline bias, then reversing the positive/negative ROIs would consequently reverse the bias, predicting a reversal of the algorithm's outcome. Of the 10 successful participants, 4 belonged to the reversal condition group. To see the results of this reversal, we looked at the four reversal participants only, and analyzed separately the first 3 d before the reversal, and the last 2 d after the reversal occurred. The results are shown in Fig. 3*C*. There was a drop in performance following the reversal, but not a reversal in outcome, which is what would be expected if these results were the consequence of a random baseline bias, rather than a result of the NF. Although the average NF performance was still slightly negative, the same trend of an increase between pre-NF rest, NF, and post-NF rest was evident.

Looking only at the reversal condition revealed that, even for just this subset of successful participants during the reversal days, the post-NF rest outcome was also significantly greater than pre-NF rest outcome ($P = 0.033$, permutation test), although the NF was not significantly greater than pre-NF rest ($P < 0.11$). Although there was a positive trend in changes between days, both between rest sessions and between NF sessions, this trend did not reach significance.

Resting-State Functional Connectivity Training Effects. To assess the daily changes in functional connectivity in the resting-state sessions before and following the NF, we first calculated, for each voxel, the changes in global connectivity. Global connectivity was defined as the average correlation of that voxel to every other voxel. This measure has previously been demonstrated to provide a sensitive marker for mapping NF training effects (24). For each participant, we subtracted the global connectivity value for each voxel in the pre-NF rest sessions from the post-NF rest sessions (averaged across days), and then performed a *t* test across participants (a form of random-effects analysis). The advantage of this approach is that it does not require any prior assumption about the training ROIs. To avoid any possible bias due to the selection of participants, we included all participants in this analysis, not just the successful ones. The results of this analysis are displayed in Fig. 4. As can be seen, there was a significant change between pre- and post-NF rest sessions in only a small number of voxels, most of which were located in left PPA, which was anatomically the most stable ROI across participants (see *Inset* in Fig. 4, showing the overlay of the PPA ROI for all successful participants). There was another small cluster of voxels in auditory cortex showing a significant increase in global connectivity, as well as in the right inferior parietal lobule. We reran this analysis for the PPA and FFA groups separately, and received a very similar map for the PPA-only group. For the FFA group, we found no significant changes between the pretraining and posttraining rest sessions that survived correction for multiple comparisons. This is possibly due to anatomically incongruent location of the FFA ROIs between participants, which makes such a group analysis difficult.

We next examined the effects of the NF on the functional connectivity to our two ROIs. To this end, we ran a functional connectivity analysis, which calculates for each voxel its degree of correlation with the good (positively rewarded) ROI, and with the bad (negatively rewarded) ROI, which for each participant was dependent on their assignment to either the FFA or the PPA group. We chose to focus on the first 3 d, which were unaffected by the reversal analysis. For each participant, the functional connectivity of each voxel to the good and bad ROIs was calculated for the average of the five NF sessions for each day. Fig. 5 shows the results of a *t*-test analysis across all 10 successful participants per day. Note that there were almost equal instances where the FFA was the good ROI as when the PPA was the good ROI. As can be clearly discerned from the figure, a preference for the good ROI started emerging on the second day, with many voxels significantly more correlated to the good ROI than to the

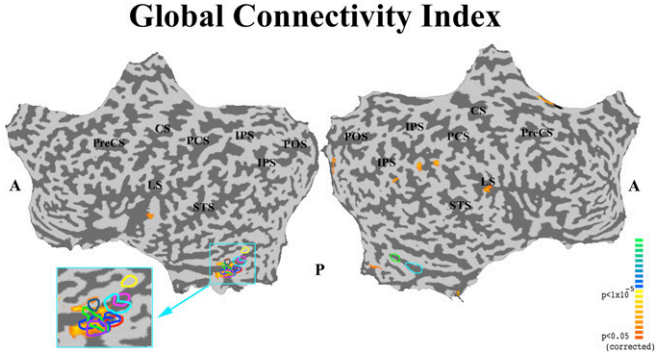

Fig. 4. Global connectivity index. The global connectivity index (the average correlation of that voxel with all other voxels) was calculated for each voxel for the pre-NF rest condition of each day. This was then subtracted from the global connectivity index for the post-NF rest condition for each day, and averaged across days. A *t* test between all subjects was then calculated ($n = 16$). The map shows all significant voxels of this analysis, thresholded with Monte Carlo correction for multiple comparisons. Colorful outlines denote the PPA ROI for each successful subject. Note that the significant voxels, found primarily in the PPA and auditory cortex, signify increased global connectivity for the post-NF rest. Abbreviations are the same as in Fig. 2.

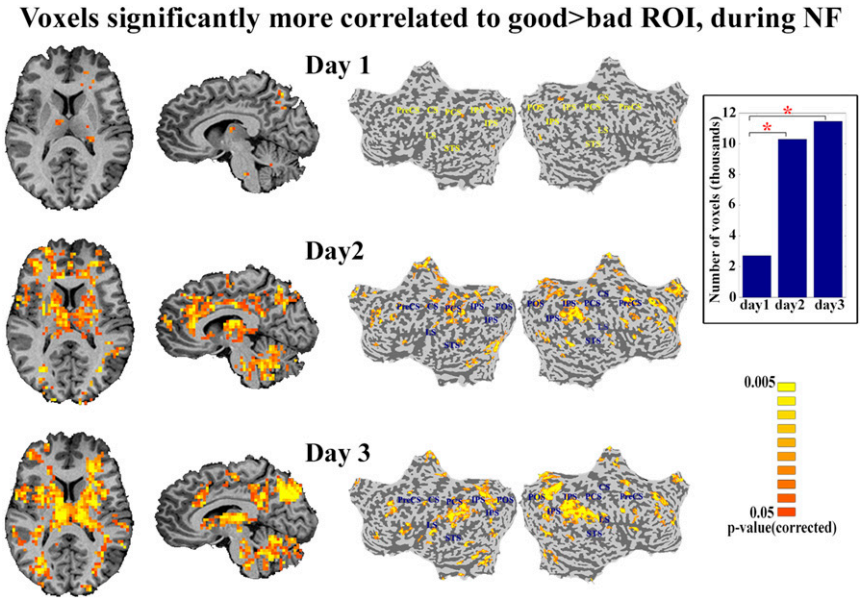

Fig. 5. Network changes in functional connectivity during the NF. Maps show voxels significantly more correlated to the good ROI than to the bad ROI, for the average NF scans on days 1–3 (*t* test across subjects, successful participants only, $n = 10$). Starting on the second day, many voxels show a preference for the good ROI, primarily in the default-mode network, cingulate cortex, thalamus, striatum, brainstem, and cerebellum. *Inset* shows the number of voxels significantly more correlated to the good ROI. A permutation test shows that the difference in this number between day 1 and day 2, as well as between day 1 and day 3, is significant. Abbreviations are the same as in Fig. 2.

bad ROI, including regions of the "default-mode" network, striatum, brainstem, and thalamus. The difference between day 2 and day 1 in the number of voxels preferentially correlated to the good vs. bad ROI was significant, as was the difference between the first and third day ($P = 0.047$, $P = 0.022$, corrected, calculated by permutation test; Fig. 5, *Inset*). An analysis of the entire group of participants shows similar but weaker results. When looking at only the first NF session of day 1, we found no voxels showing a significant preference to the good ROI.

Discussion

Spontaneous Connectivity Changes Following Covert Activity–Reward Associations. Our results are a further indication that rtfMRI-based NF can succeed without an explicit strategy, as has previously been shown (8, 16). However, the present study takes the implicit learning paradigm described in those papers a step further, by removing all awareness to the association between reward and brain activations. The successful participants showed a link between the modulation of the trained networks during the NF, and subsequent changes in resting-state connectivity, without being aware that activity–reward association was implemented in their brain, and without intending or attempting to learn. Furthermore, they did so without any knowledge of having this reward–activity association implemented, as shown by the postexperimental questionnaires. Finally, our results expand the range of training strategies by demonstrating that a NF effect can be achieved using localized univariate signals as targets rather than multivariate patterns used in previous research (16).

An important issue concerns the possibility that the results reported here were due to effects unrelated to the NF. In particular, slow random fluctuations in BOLD signal and correlations as well as random structural biases should be ruled out.

The random division into FFA and PPA groups was intended to rule out any structural bias in favor of one of those ROIs that might influence the algorithm's output, while creating an intrinsic control group. The distribution of successful participants between the two groups (six FFA, four PPA) is indicative that there was no such consistent bias toward one ROI or the other that could explain the observed NF learning. It is possible that, although there was no consistent structural bias in favor of either FFA or PPA, each participant had a small individual bias in favor of one of these ROIs. The reversal procedure was used to control for such an individual baseline bias, by testing both FFA-positive and PPA-positive directions of the algorithm on the same participant. Four of six participants were successful in recovering from this reversal, and although there was a decline in performance, which might be expected after 3 d of training in the opposite direction, there was no reversal of the results, which is what would be expected if the algorithm output were a reflection of a simple baseline bias (Fig. 3*C*).

There remains the issue of the slow fluctuations, which might in theory account for the link we see between the pre-NF rest, NF, and post-NF sessions. Because it is difficult to distinguish such putative "slow fluctuations" from real changes, the best way to attempt to differentiate the two was to look at the difference between the successful and the unsuccessful participants in this regard. Fig. 3 *A* and *B* clearly shows that, although there was a clear link (i.e., a significant correlation between the NF success and the improvement between rest sessions for each day) for the successful participants, such a trend was absent among the unsuccessful participants. This was further verified in the permutation tests, which found the difference between the post-NF rest/pre-NF rest to be significant for the successful subjects, but not for the unsuccessful ones.

A related question is why some participants did not show NF-related changes. It should be noted that the task was highly demanding, requiring differentiation of two neighboring visual areas, and participants were hindered by many factors. Note also that, because of the hemodynamic delay, the feedback was delayed by ~6 s. Although it has been shown that different participants can have slightly different hemodynamic delays (28), ranging from as low as 4 s, this should not influence the algorithm output, as the algorithm itself does not take this delay into account. Rather, it simply applies the rule for the signal measured each TR, regardless of when that neuronal activity was generated. This would mean that the feedback delay for different

participants might be slightly different, although consistent for each participant. Moreover, many other NF studies with simpler learning tasks, as well as EEG studies that do not suffer from built-in delay, also report that a certain percentage of participants fail the task for unknown reasons (6, 10, 29, 30). This in fact is such a widespread phenomenon in NF and brain–computer interface (BCI) literature that it has its own term, BCI illiteracy (31, 32).

Another possibility that merits further exploration, given our current data showing global connectivity changes involving PPA but not FFA (Fig. 4), is that some cortical areas may be easier to train than others. Future studies should further test this hypothesis.

To ensure that the feedback itself did not somehow influence the results, we decided to deliver it in a different modality. Although most NF studies of sensory and particularly visual areas use same modality feedback (8, 16), other studies on higher-order cognitive ROIs have shown that such cross-modal integration of feedback is possible (7, 24). Such cross-modal integration of feedback is in itself an interesting phenomenon, suggesting an integrated reward-learning network (33).

Networks Involved. Fig. 4 shows that most of the consistent changes in global connectivity levels, which persisted beyond the NF sessions themselves and into the post-NF rest session, were found in one of the task ROIs, left PPA. There were also some traces in another strongly task-related region, the auditory cortex. This increase in global activation in PPA is likely related to the large network that could be seen to coactivate along with the good ROI (Fig. S1). Because the individual ROIs were defined according to an independent functional localizer, there was not much overlap between the FFA ROIs of individual participants, which might explain why we did not see this rise in global connectivity in FFA on the average of all participants. Alternatively, there may be some inherent differences in the relative sensitivity of the FFA to NF manipulations. We have recently reported that the PPA is significantly more sensitive to NF-fMRI activation compared with the FFA (34).

The overlap in the individual PPA ROIs, however, was greater, and the changes in global connectivity were indeed found where the ROIs from the greatest number of participants overlapped (Fig. 4, *Inset*).

Fig. 4 provides evidence that the changes in global connectivity were well localized to NF-related areas. As can be seen from Fig. 5, the networks involved in the increased connectivity to the good ROI were widespread, and included areas known to be involved in learning and reward, such as the brainstem, lentiform nucleus, and the caudate (35–37). Areas of the default-mode network were also significantly more correlated to the good ROI, indicating a possible role of this network in the NF effect. Because the identity of the good and bad ROIs was counterbalanced between FFA/PPA, this cannot be a structural change related to one of those regions. Note that these changes in functional connectivity are not predicted by success as measured by the algorithm, which is triggered by greater activity in the good vs. the bad ROI.

The widespread nature of the connectivity changes is compatible with previous work showing that NF training may affect entire networks. For example, targeted NF activation of the anterior cingulate cortex led to a long-term change in a widespread frontoparietal network (24).

In the present study, we failed to find a long-term improvement in performance across days. A number of factors may account for this. A major limitation is the long hemodynamic delay that induced a long (4–6 s) temporal gap between the neuronal activity and the reward. Additional potential problems could be the limited number of training events and the long intersession interruptions in which competing associations may have weakened the training trace.

It should be noted that a transient trace of the training did persist in the connectivity structure measured in the rest session immediately following the training (Fig. 4). Although this may seem surprising, a growing body of recent research points to such resting-state connectivity changes as a sensitive marker for prior training and individual experience. In fact, we have previously proposed that traces of cortical networks' coactivations during a NF task or during habitual cortical activation could later be seen in enhanced connectivity during the resting state (38, 39). Fig. 5 itself also shows long-term changes in connectivity related to our two training ROIs. These results are in line with the well-documented effect of NF training on changes in network functional connectivity (24, 40–43). The changes in resting-state connectivity following training thus extend previous NF training based on spontaneous fluctuations that showed changes in activation during training (16, 44).

Increased Activation in High-Order Areas Does Not Necessarily Generate Conscious Percepts. The present findings contribute insights into another fundamental question, which is the extent to which we are aware of the slow spontaneous fluctuations that emerge in cortical networks. Although there were widespread changes both in connectivity (Figs. 4 and 5) and in overall activation levels associated with the NF (Fig. S1), participants did not make the connection between any conscious percepts (thoughts, imagery) they experienced during the NF, and the feedback they received. Nor were they able, given a forced-choice questionnaire, to correctly guess which areas were responsible for the feedback. This failure reveals that participants did not become aware of their NF-related spontaneous fluctuations. These results are important because they extend the previously reported findings of failure to become aware of specific V1 activation patterns (16), to high-order visual areas, whose activity has been consistently demonstrated to be more closely linked to perceptual awareness (45, 46). Together, these findings argue against the suggestion that the slow spontaneous fluctuations reflect stream of conscious thoughts and images, but rather that this activity remains largely subliminal (47–49).

Conclusions and Implications. Essentially, NF can be considered a way of teaching participants to control cortical function by creating a new feedback pathway, or in a sense, establishing a new reward-control loop. The present study demonstrates that such a reward-control loop can induce connectivity changes even without the participant's knowledge and awareness.

These new control loops are the true power of NF. They could potentially allow us to train and correct widespread network configurations in the brain that are associated with various pathologies, which might be difficult to control via an explicit task or strategy. Thus, an implicit NF, such as reported here, might represent a major advance in our ability to manipulate such networks. The covert NF paradigm described in this paper, when further optimized, may have potential uses with severe clinical populations, for which task compliance, and even conscious awareness is sometimes difficult to ensure. Further research is necessary to optimize this implicit NF methods in other conditions as well.

Methods

Participants. Eighteen healthy participants (11 women, aged 24–35) participated in the experiments. Two participants (1 woman) were disqualified after the first day, due to excessive motion during the scans (>1 mm). The remaining 16 participants participated in all 5 d of the experiment. All participants were right-handed, and had normal or corrected-to-normal vision. The Tel-Aviv Souraski Medical Center ethics committee approved the protocol and informed consent was obtained from all participants.

Experiments.

Visual localizer experiment. A conventional static object-visual localizer was used to identify FFA and PPA in each subject (50). Data from such a localizer was either already available for participants who had previously participated in prior studies in our laboratory, or was collected between 2 wk to 24 h before the beginning of the NF experiment, by a different experimenter, ostensibly for the purpose of a completely different study. On post hoc questioning, not a single subject suspected any connection between the visual localizer experiment and the NF experiment.

NF experiment. Participants were scanned for 5 d. These days were consecutive whenever possible (consecutive for 13 of 16 participants, conducted over the course of 7 d for the remaining 3 of 16 participants). Each day consisted of an anatomical scan, followed by seven functional scans, in this order: a 9-min rest scan, five iterations of a 10-min NF scan, and a final 9-min rest scan. All scans were conducted in total darkness, with eyes closed. The anatomical scan was conducted first, to allow the coregistration of scans on different days to the same anatomical space. Participants were randomly assigned to either the FFA-positive or PPA-positive group. Participants were randomly assigned to either use their right hand for good responses and left for bad, or vice versa. For visualization purposes in the figures, auditory cortex was defined based on a random-effects group analysis of the feedback sounds of the first NF session of the first day for each subject (all sounds > no sound). Note that this analysis uses only a small subset of the data.

NF Algorithm. The NF algorithm looked at differences in activation between the two target ROIs. The raw average activation value at each TR was first normalized for each ROI individually. To achieve this, the first 15 TRs were collected as baseline, and no feedback was provided. The median value during this time period was considered the baseline. This baseline was continuously updated for the first 50 TRs as more data were collected, and after the first 50 TRs, a sliding window was used so that the baseline was always calculated as the median value over the last 50 TRs. The sliding window was implemented to compensate for any drift in the signal. For each TR, good feedback was provided if the normalized activation in the target good ROI (for that TR), divided by the normalized activation in the target bad ROI (for that TR), was above a certain threshold, as well as above the baseline activation for the good ROI (i.e., the median value of the last 50 TRs). This rule was for each single TR separately; there was no demand for the signal to be higher than threshold for more than one TR. As a result of this rule, good feedback was only ever awarded if the activation levels in the good ROI could be said to be positive, and higher than activation levels in the bad ROI. Bad feedback was provided if the normalized activation in the bad ROI divided by the normalized activation in the good ROI was above the same threshold, and also the activation in the bad ROI was higher than the baseline of that ROI. The threshold was chosen so that either positive or negative feedback would be received for roughly one-third of the TRs. This threshold was calculated by running a simulation of the algorithm on previously collected rest data. The identity of the good ROI (whether FFA or PPA) was randomly decided for each participant. For reversal participants, this identity was flipped on the beginning of the fourth day.

Postscan Interviews. Immediately following the last scan session on the last day, participants underwent a detailed interview focused on assessing their awareness of the study purpose and the NF. Participants were first asked to write down any thoughts they had on the purpose of the experiment, and any thoughts/feelings associated or elicited by the feedback sounds. Next, participants were asked whether they thought they had any influence over the sounds they had heard during the experiment. Finally, the purpose of the experiment as a NF study relying on two areas of cortex was revealed, and participants were asked whether they had any notion of what might be driving the feedback. All participants reported having no notion at all as to what this might be. Reversal subjects were also asked whether they felt any change in the feedback at some point during the experiment, to which they all replied negatively. All participants were then given a five-alternative forced choice for which region/function was associated with the positive and negative rewards: face-related imagery, places/houses-related imagery, abstract visual forms, language, and body/motion-related imagery, resulting in guesses at chance level.

Imaging Setup. The scans were performed on a 3-T Trio Magnetom Siemens scanner at the Weizmann Institute of Science (Rehovot, Israel). Three-dimensional T1-weighted anatomical images were acquired with high-resolution 1-mm slice thickness [3D magnetization-prepared rapid acquisition with gradient echo sequence; TR, 2,300 ms; echo time (TE), 2.98 ms; 1 × 1 × 1-mm voxels]. BOLD contrast was obtained with gradient echo-planar imaging sequence [TR, 2,000 ms; TE, 30 ms; matrix size, 80 × 80; scanned volume, 32 axial slices of 3-mm thickness (no gap, 3 × 3 × 4-mm voxel), anterior commissure/posterior commissure].

Data Analysis and Preprocessing. rtfMRI data were analyzed with the "Turbo-BrainVoyager" (TBV) software package, a real-time processing, analysis, and visualization application, which receives dicoms from the scanner, along with complementary in-house software. A feature of TBV allows coregistration of scans on different days, thus ensuring anatomical accuracy for the selected target ROIs. Preliminary data preprocessing such as motion correction was carried out in TBV, and then the average raw values of each ROI were saved for each TR. These values were then read by our in-house software using Matlab, which executed the NF algorithm for each time point (see above). The algorithm then determined the appropriate feedback for each time point (positive, negative, or none), which was then delivered to the participant through MRI-compatible headphones (MR Confon). Participant button presses were recorded using the Matlab PsychToolbox.

fMRI data were analyzed with the "BrainVoyager" software package (Brain Innovation) and with complementary in-house software. The first two images of each functional scan were discarded. The functional images were superimposed on 2D anatomic images and incorporated into the 3D datasets through trilinear interpolation. The cortical surface in a Talairach coordinate system (51) was reconstructed for each subject from the 3D-spoiled gradient echo scan. Preprocessing of functional scans included 3D motion correction and filtering out of low frequencies up to three cycles per scan (slow drift). Statistical analysis/mapping was based on the GLM, with the regressor built on events of positive/negative feedback either without applying the standard hemodynamic response function (equivalent to shifting back 6 s in time) for the results shown in Fig. S1, or unshifted in time for the results in Fig. 2. The analysis was performed independently for the time course of each individual voxel. After computing the coefficients for the regressor, Student's *t* test was performed. In calculating *P* values, the autoregression factor was taken into account (BrainVoyager software package), because consecutive fMRI data points of the regressor are not statistically independent due to the nature of the hemodynamic response. Student's *t* test for each voxel between subjects was then carried out. The multisubject functional maps were projected on an unfolded Talairach normalized brain. Significance levels were calculated, taking into account the minimum cluster size and the probability threshold of a false detection of any given cluster. This was accomplished by a Monte Carlo simulation (AlphaSim by B. Douglas Ward, Medical College of Wisconsin, Milwaukee), using the combination of individual voxel probability thresholding and minimum cluster size of 30–52 voxels; the probability of a false-positive detection per image was determined from the frequency count of cluster sizes within the entire cortical surface.

ACKNOWLEDGMENTS. We thank the participants for volunteering to take part in the study.

1. Birbaumer N (2006) Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology* 43(6):517–532.
2. Felton EA, Wilson JA, Williams JC, Garell PC (2007) Electrocorticographically controlled brain-computer interfaces using motor and sensory imagery in patients with temporary subdural electrode implants. Report of four cases. *J Neurosurg* 106(3):495–500.
3. Hui M, Zhang H, Ge R, Yao L, Long Z (2014) Modulation of functional network with real-time fMRI feedback training of right premotor cortex activity. *Neuropsychologia* 62:111–123.
4. Cohen O, Koppel M, Malach R, Friedman D (2014) Controlling an avatar by thought using real-time fMRI. *J Neural Eng* 11(3):035006.
5. Bagdasaryan J, Quyen MleV (2013) Experiencing your brain: Neurofeedback as a new bridge between neuroscience and phenomenology. *Front Hum Neurosci* 7:680.
6. Robineau F, et al. (2014) Self-regulation of inter-hemispheric visual cortex balance through real-time fMRI neurofeedback training. *Neuroimage* 100:1–14.
7. Seitz AR (2013) Cognitive neuroscience: Targeting neuroplasticity with neural decoding and biofeedback. *Curr Biol* 23(5):R210–R212.
8. Scharnowski F, Hutton C, Josephs O, Weiskopf N, Rees G (2012) Improving visual perception through neurofeedback. *J Neurosci* 32(49):17830–17841.
9. deBettencourt MT, Cohen JD, Lee RF, Norman KA, Turk-Browne NB (2015) Closed-loop training of attention with real-time brain imaging. *Nat Neurosci* 18(3):470–475.
10. deCharms RC, et al. (2005) Control over brain activation and pain learned by using real-time functional MRI. *Proc Natl Acad Sci USA* 102(51):18626–18631.
11. Yuan H, et al. (2014) Resting-state functional connectivity modulation and sustained changes after real-time functional magnetic resonance imaging neurofeedback training in depression. *Brain Connect* 4(9):690–701.
12. Grone M, et al. (2015) Upregulation of the rostral anterior cingulate cortex can alter the perception of emotions: fMRI-based neurofeedback at 3 and 7 T. *Brain Topogr* 28(2):197–207.

13. Lawrence EJ, et al. (2013) Self-regulation of the anterior insula: Reinforcement learning using real-time fMRI neurofeedback. *Neuroimage* 88C:113–124.
14. Weiskopf N, et al. (2007) Real-time functional magnetic resonance imaging: Methods and applications. *Magn Reson Imaging* 25(6):989–1003.
15. Koush Y, et al. (2013) Connectivity-based neurofeedback: Dynamic causal modeling for real-time fMRI. *Neuroimage* 81:422–430.
16. Shibata K, Watanabe T, Sasaki Y, Kawato M (2011) Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* 334(6061): 1413–1415.
17. Reber AS (1989) Implicit learning and tacit knowledge. *J Exp Psychol Gen* 118(3): 219–235.
18. Wulf G, Weigelt C (1997) Instructions about physical principles in learning a complex motor skill: To tell or not to tell..... *Res Q Exerc Sport* 68(4):362–367.
19. Adams JA (1971) A closed-loop theory of motor learning. *J Mot Behav* 3(2):111–149.
20. Gomi H, Kawato M (1993) Neural-network control for a closed-loop system using feedback-error-learning. *Neural Netw* 6(7):933–946.
21. Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci* 17(11):4302–4311.
22. Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. *Nature* 392(6676):598–601.
23. Hutchison RM, Culham JC, Everling S, Flanagan JR, Gallivan JP (2014) Distinct and distributed functional connectivity patterns across cortex reflect the domain-specific constraints of object, face, scene, body, and tool category-selective modules in the ventral visual pathway. *Neuroimage* 96:216–236.
24. Harmelech T, Preminger S, Wertman E, Malach R (2013) The day-after effect: Long term, Hebbian-like restructuring of resting-state fMRI patterns induced by a single epoch of cortical activation. *J Neurosci* 33(22):9488–9497.
25. Taubert M, Lohmann G, Margulies DS, Villringer A, Ragert P (2011) Long-term effects of motor training on resting-state networks and underlying brain structure. *Neuroimage* 57(4):1492–1498.
26. Tambini A, Ketz N, Davachi L (2010) Enhanced brain correlations during rest are related to memory for recent experiences. *Neuron* 65(2):280–290.
27. Lewis CM, Baldassarre A, Committeri G, Romani GL, Corbetta M (2009) Learning sculpts the spontaneous activity of the resting human brain. *Proc Natl Acad Sci USA* 106(41):17558–17563.
28. Aguirre GK, Zarahn E, D'esposito M (1998) The variability of human, BOLD hemodynamic responses. *Neuroimage* 8(4):360–369.
29. Bray S, Shimojo S, O'Doherty JP (2007) Direct instrumental conditioning of neural activity using functional magnetic resonance imaging-derived reward feedback. *J Neurosci* 27(28):7498–7507.
30. Chiew M, LaConte SM, Graham SJ (2012) Investigation of fMRI neurofeedback of differential primary motor cortex activity using kinesthetic motor imagery. *Neuroimage* 61(1):21–31.
31. Allison BZ, Neuper C (2010) Could anyone use a BCI? *Brain–Computer Interfaces*, eds Tan DS, Nijholt A (Springer, London), pp 35–54.
32. Vidaurre C, Blankertz B (2010) Towards a cure for BCI illiteracy. *Brain Topogr* 23(2): 194–198.
33. Fuster JM, Bodner M, Kroger JK (2000) Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature* 405(6784):347–351.
34. Harmelech T, Friedman D, Malach R (2015) Differential magnetic resonance neurofeedback modulations across extrinsic (visual) and intrinsic (default-mode) nodes of the human cortex. *J Neurosci* 35(6):2588–2595.
35. Delgado MR, Stenger VA, Fiez JA (2004) Motivation-dependent responses in the human caudate nucleus. *Cereb Cortex* 14(9):1022–1030.
36. Berridge KC, Robinson TE (1998) What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Res Brain Res Rev* 28(3):309–369.
37. Haber SN, Knutson B (2010) The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology* 35(1):4–26.
38. Harmelech T, Malach R (2013) Neurocognitive biases and the patterns of spontaneous correlations in the human cortex. *Trends Cogn Sci* 17(12):606–615.
39. Ramot M, et al. (2013) Emergence of sensory patterns during sleep highlights differential dynamics of REM and non-REM sleep stages. *J Neurosci* 33(37):14715–14728.
40. Ros T, et al. (2013) Mind over chatter: Plastic up-regulation of the fMRI salience network directly after EEG neurofeedback. *Neuroimage* 65:324–335.
41. Haller S, et al. (2013) Dynamic reconfiguration of human brain functional networks through neurofeedback. *Neuroimage* 81:243–252.
42. Scheinost D, et al. (2013) Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity. *Transl Psychiatry* 3:e250.
43. Zotev V, et al. (2011) Self-regulation of amygdala activation using real-time FMRI neurofeedback. *PLoS One* 6(9):e24522.
44. Scharnowski F, et al. (2014) Connectivity changes underlying neurofeedback training of visual cortex activity. *PLoS One* 9(3):e91090.
45. Fisch L, et al. (2009) Neural "ignition": Enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron* 64(4):562–574.
46. Grill-Spector K, Malach R (2004) The human visual cortex. *Annu Rev Neurosci* 27: 649–677.
47. Yellin D, Berkovich-Ohana A, Malach R (2015) Coupling between pupil fluctuations and resting-state fMRI uncovers a slow build-up of antagonistic responses in the human cortex. *Neuroimage* 106:414–427.
48. Ramot M, et al. (2011) Coupling between spontaneous (resting state) fMRI fluctuations and human oculo-motor activity. *Neuroimage* 58(1):213–225.
49. Moutard C, Dehaene S, Malach R (2015) Spontaneous fluctuations and non-linear ignitions: Two dynamic faces of cortical recurrent loops. *Neuron* 88(1):194–206.
50. Hasson U, Harel M, Levy I, Malach R (2003) Large-scale mirror-symmetry organization of human occipito-temporal object areas. *Neuron* 37(6):1027–1041.
51. Talairach J, Tournoux P (1988) *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System: An Approach to Cerebral Imaging* (Georg Thieme, Stuttgart), p 122.

Supporting Information

Ramot et al. 10.1073/pnas.1516857113

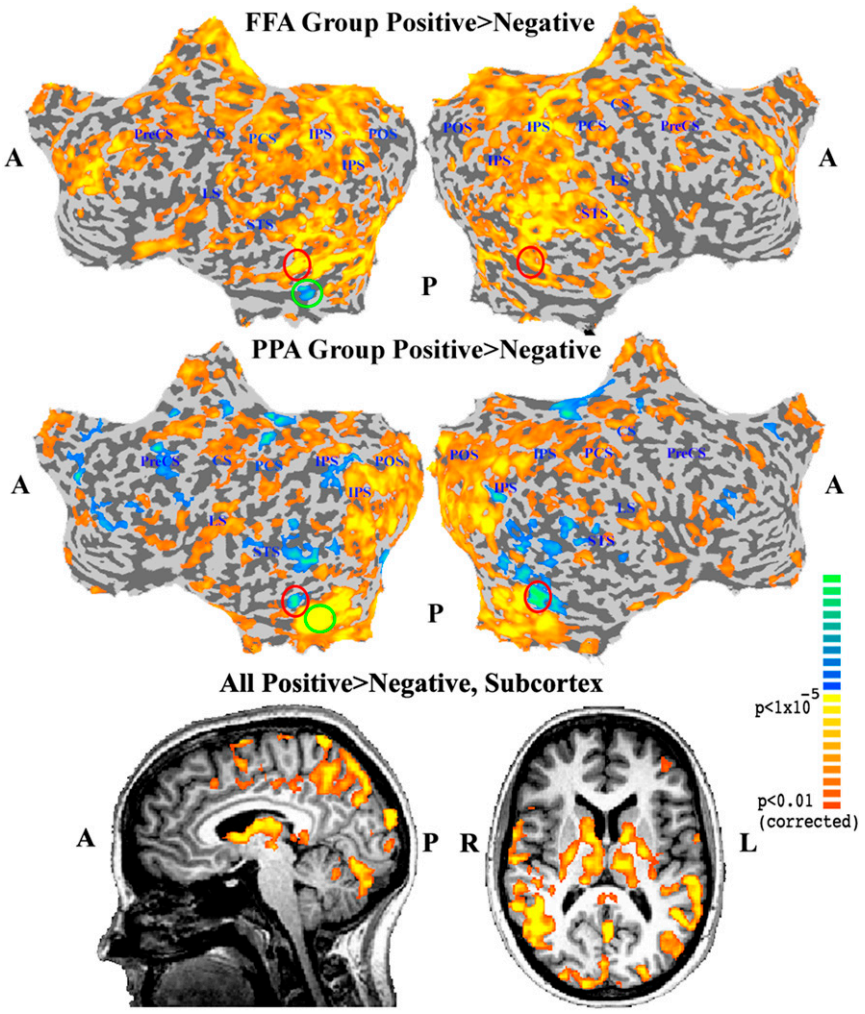

Fig. S1. Mapping the algorithm onto the brain. A protocol based on the algorithm output was constructed for the NF sessions, but the standard hemodynamic response function (HRF), which essentially shifts the predictor 6 s in time, was not applied. The resulting GLM thus reflects the underlying neuronal events that precipitated the algorithm's outcome, rather than the feedback events. The top panel shows the positive > negative contrast for the FFA group participants, the middle panel shows the same for the PPA group, and the bottom panel shows the results of the positive > negative contrast for both groups, in the subcortex. Note that, as expected, the FFA group shows higher activation in FFA and decreased activation in PPA, whereas the PPA group shows the opposite. Approximate group locations of FFA and PPA ROIs are marked with red (FFA) and green (PPA). L, left; R, right; all other abbreviations are the same as in Fig. 2.