bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: Direct modulation of aberrant brain network connectivity through real-time NeuroFeedback

Authors: Michal Ramot[1,*], Sara Kimmich[1], Javier Gonzalez-Castillo[1], Vinai Roopchansingh[2], Haroon Popal[1], Emily White[1], Stephen J. Gotts[1], Alex Martin[1]

Affiliations:

[1]Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892

[2]Functional MRI Facility, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892

*Correspondence to: michal.ramot@nih.gov

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The existence of abnormal connectivity patterns between resting state networks in neuropsychiatric disorders, including Autism Spectrum Disorder (ASD), has been well established. Traditional treatment methods in ASD are limited, and do not address the aberrant network structure. Using real-time fMRI neurofeedback, we directly trained 3 brain nodes in participants with ASD, in which the aberrant connectivity has been shown to correlate with symptom severity. 17 ASD participants and 10 control participants were scanned over multiple sessions (123 sessions in total). Desired network connectivity patterns were reinforced in real-time, without participants' awareness of the training taking place. This training regimen produced large, significant long-term changes in correlations at the network level, and whole brain analysis revealed that the greatest changes were focused on the areas being trained. These changes were not found in the control group. Moreover, changes in ASD resting state connectivity following the training were correlated to changes in behavior, suggesting that neurofeedback can be used to directly alter complex, clinically relevant network connectivity patterns.

Significance Statement

Many disorders are characterized by underlying abnormalities in network connectivity. These abnormalities are difficult to address with explicit training procedures (which are unlikely to target the specific abnormalities). Covert neurofeedback however, can directly target these networks, positively reinforcing the desired connections. We have developed a method for reinforcing correlations in real-time, and show that such training is effective, inducing significant, long-lasting changes in connectivity between aberrant networks in Autism Spectrum Disorder. This provides a potential mechanism for modulating aberrant correlation structures in other clinical groups as well.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Autism Spectrum Disorder (ASD) refers to a group of neurobiological disorders, which affect a growing proportion of the population. Patients with ASD suffer from a range of social and communication impairments, along with other characteristic behaviours and deficits. Behavioural treatment options are limited in their efficacy, and often do not generalize well beyond the specific training paradigm (Otero et al., 2015, Williams White et al., 2007). Numerous studies have documented widespread patterns of aberrant brain functional connectivity in patients with ASD, involving many cortical regions including frontal, parietal, and temporal lobes (Muller et al., 2011, Picci et al., 2016a, Di Martino et al., 2014). Specifically, these studies show that multiple cortical areas are significantly under-connected in patients with ASD compared to typically developing (TD) control subjects, although over-connectivity has also been reported (Hahamy et al., 2015, Belmonte et al., 2004). At the individual level, the degree of diminished connectivity as well as measures of cortical thickness were found to be correlated with symptom severity (Gotts et al., 2012, Wallace et al., 2012) and this measure of connectivity was even predictive of future progression of autistic symptoms (Plitt et al., 2015). Together, these findings suggest a causal link between connectivity and behavior, such that changing the connectivity might lead to a change in behavior. Traditional, explicit training paradigms, activate these aberrant networks (Kana et al., 2009, Di Martino et al., 2009a), thus potentially reinforcing the sub-optimal connectivity. An implicit training paradigm, which allows the direct reinforcement of the desired networks, bypassing the atypical activation induced by explicit tasks, might be a better candidate for potential intervention in ASD.

Real-time fMRI neurofeedback (rt-fMRI-nf) is an emerging technique with great potential for clinical applications (Stoeckel et al., 2014, Sulzer et al., 2013, Weiskopf, 2012, Birbaumer et al., 2013). With this technique, network states can be monitored in real-time, and desired states can be reinforced through positive feedback. Covert neurofeedback is a variant of neurofeedback, in which participants are given no strategy with which to control the feedback, and might not even be cognizant that the feedback is related to brain activity. This tool is therefore extremely flexible, as it does not require the formulation of a specific strategy and is not limited by what we know about the networks or the ways in which they are typically activated. Instead, desired states are reinforced when they occur spontaneously, allowing for implicit training of networks, such as

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

those found to be under-connected in ASD. We designed a covert neurofeedback experiment, to test whether it would be possible to change connectivity between these aberrantly connected network nodes, through direct reinforcement of spontaneously occurring network states. This decision is motivated by recent work showing that positive and negative reinforcement of brain activity patterns are sufficient for promoting small but widespread changes in network connectivity, even without any learning intention on the part of participants (Ramot et al., 2016).

Further evidence that covert neurofeedback can change networks, and that these changes can have behavioral effects, comes from other recent work. A number of studies have been successful at training complex patterns of activity within a given network using multi-voxel pattern analysis (MVPA) techniques (deBettencourt et al., 2015, Amano et al., 2016, Shibata et al., 2011), with feedback related changes corresponding to robust behavioural changes after only a few sessions, and significant change being detectable after as little as one training session. Feedback induced behavioural changes have been shown to range from very local and low level, such as changes in perception of line orientation after V1 training (Shibata et al., 2011), or inducing color association in V1 in a somewhat less local paradigm (Amano et al., 2016), to changes in high level functions such as attention (deBettencourt et al., 2015), and fear perception (Koizumi et al., 2016). Such changes can even be bi-directional, both behaviourally and at the network level (Cortese et al., 2017). This previous work sets up covert neurofeedback as a good candidate for a potential intervention in ASD, though whether specific, long-ranging connectivity changes can be induced through neurofeedback, which regions / networks are amenable to such reinforcement, how such training will affect wider brain networks, and how long these changes last, are all still open questions. In this study, we lay out a proof of principle of the plausibility of such training, showing robust, long-lasting feedback induced changes in these aberrant networks, coupled with preliminary results as to the behavioural correlates of these changes. These results reflect not only on the potential uses of such training in ASD, but also in other disorders with underlying aberrant connectivity at their core.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Selection of training targets

We used previously collected resting state data on large groups of ASD and TD participants (N = 56 ASD, 62 TD) to identify two target brain regions that showed large under-connectivity in ASD compared with TD individuals, while also being physically distant from each other, and belonging to separate networks (Fig. 1): target1 in superior temporal sulcus (STS) and target2 in somatosensory cortex, both of which have been consistently implicated in social processing (Allison et al., 2000, Frith and Frith, 2010, Adolphs, 2009, Damasio et al., 2000), and have previously been found to be under-connected and atypically activated in ASD (Chen et al., 2015, Gotts et al., 2012, Müller et al., 2001, Tuttle et al., 2016, Khan et al., 2015, Khan et al., 2013). This dataset is an expansion of previously reported data (Gotts et al., 2012), which found very similar aberrant connectivity patterns, matching results from other studies using large datasets (Cheng et al., 2015). As in the previously published subset of the dataset, we found that under-connectivity between these two networks (STS and somatosensory) in ASD was significant in both this large dataset and in the 17 participants recruited for the neurofeedback study ($p < 0.001$ for both cohorts), as well as significantly correlated to social symptom severity, as measured by the Social Responsiveness Scale (SRS) ($r= -0.35$ $p<0.009$ without regressors, $r = -0.31$, $p < 0.026$, using age and motion as regressors). The SRS is a parent filled questionnaire, which is designed to be a continuous, cardinal measure of social symptom severity in ASD, and has been shown to correlate with functional brain connectivity measures in multiple studies (Anderson et al., 2011, Di Martino et al., 2009b). This result indicates that connectivity between these two networks is clinically relevant, i.e. the lower the connectivity, the more severe the social symptoms (higher score on the SRS). The first goal of the training was therefore to increase the connectivity between target1 in STS, and target2 in somatosensory cortex.

In order to ascertain that we would only be reinforcing connectivity between our two targets, rather than global changes that cause an overall increase in correlations across the entire brain in an undifferentiated manner, we selected a third control region (in the inferior parietal lobule or IPL, part of the default mode network), which was chosen for being uncorrelated to the two target regions in our dataset of TD participants during resting state. IPL was significantly over-correlated to STS target1 in the ASD cohort participating in this study (Fig. 1C). This

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

combination of under-connectivity between STS and somatosensory with over-connectivity to the default mode network, is in line with recent evidence of reduced within-network cohesion coupled with reduced between-network differentiation (Hahamy et al., 2015, Keown et al., 2016). The goal of the neurofeedback training was therefore to induce greater differentiation between these three regions of interest (ROIs) in participants with ASD, so as to bring connectivity levels between those three networks closer to those of TD individuals. This meant increasing connectivity between the two target regions, while simultaneously decoupling the two target-control pairs. To this end, we came up with the composite difference measure, combining the target-target and target-control correlations (see Methods). This measure was also significantly different between the ASD group and the TD group in both the previous large dataset, and in our cohort ($p < 0.004$ for both cohorts, Fig 1B-C).

Training paradigm

For the initial part of the study, 17 patients with ASD participated in four training sessions, over the course of 8 days (two sessions of two consecutive days each, a week apart). Each session consisted of two rest scans, followed by four neurofeedback training scans, and finally two more rest scans (Each scan was 9 minutes in duration. See Fig. 2). During the neurofeedback scans, participants started with a blank screen, and were instructed to attempt to reveal the picture hidden underneath (see fig. S1 for an example). This was described to them as a puzzle task. No further instructions were given. Parents filled out behavioral questionnaires before training began, and two weeks after the last training session. An additional follow up study was then carried out, in which 15 of the 17 original participants returned for a final, slightly shorter training session. The interval between the original training and the follow up varied greatly between subjects, and ranged from 5 to 56 weeks.

One of the barriers to carrying out connectivity-based rt-fMRI-nf has been the slow timescale of fMRI recordings, making online calculations of correlations very limited. We therefore developed a method that can approximate the correlations using only two time points: every two seconds, for each TR (time to repetition), the signals from the three ROIs were analyzed in real-time (see Methods), and the trend in the signal compared to the previous TR was noted for each of the three ROIs (increase / decrease). Positive feedback, in the form of revealing a part of the

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

picture accompanied by an upbeat sound, was given whenever the network was deemed to have reached its desired state. As our goal was to increase correlation between the two target regions, and decrease correlations between the target and the control regions, feedback (i.e. revealing a part of the picture) was given whenever the signal trend in the two target ROIs was the same, and opposite from the trend in the control ROI (Fig. 2, Methods). This "two-point" method was validated as being a good proxy for correlation analysis by comparing the results from this to standard Pearson's correlation offline ($r = 0.61$, $p < 1x10^{-4}$ permutation test, fig. S2).

At the end of each neurofeedback scan, participants were presented with their score, i.e. how many picture squares they had managed to reveal. They were then given a chance to attempt to beat their score on the next run, to win an additional bonus on top of the normal study compensation. The pictures were chosen to be neutral, depicting mostly scenes devoid of people and text, or abstract art/objects. Random pictures were chosen for each run, from a large set of such pictures.

Participants were blind to the purpose of the study, as well as to the mechanism of the neurofeedback, and even to the fact that it was neurofeedback. This was ascertained by exit questionnaires at the end of the last day of scanning, in which participants were interviewed regarding their thoughts on the study, their motivation, and their strategies during the training (see table S1 for responses). Responses as to the perceived nature of the "puzzle task" varied widely, as did reported strategies, but none held any resemblance to the neurofeedback algorithm. Strategies mostly revolved around different ways of looking at the picture, as it was being revealed. Despite not knowing what they were supposed to do, most participants were highly motivated to solve the puzzles, with only 3 of the 17 participants reporting a motivation score of less than 5 (on a scale of 1-10, see table S1).

Control group

An additional control group of 10 TD participants completed the same initial 4-day training regimen, following the same protocol as the above. This group received feedback on the same three nodes, but in a different network configuration: target1 in the STS remained the same, but somatosensory target2 and the IPL control region switched roles, so that feedback was given whenever STS (target1) and IPL (now target2) were co-modulated, and were opposite to

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

somatosensory cortex (now control), see fig. S3. This provided feedback orthogonal to that given to the ASD group. Another key difference is that this feedback was antithetical to the normal connectivity patterns found in the typically developing brain, as STS and somatosensory are well correlated in the typically developing brain, whereas the IPL region used in this study was explicitly chosen to be as uncorrelated as possible with STS in TDs during rest (Fig. 1B-C). This control therefore served a dual purpose: in terms of the network that the ASD participants were being trained on, which rewarded increased connectivity between STS and somatosensory and decoupling of these from IPL, this was random feedback. That is to say, the feedback given to the TD participants was uncorrelated with the feedback they would have received had they been trained on the same network configuration as the ASD participants. This served as a control for any changes in connectivity in that direction being driven by something other than the feedback. At the same time, this control also examined whether it is possible to modulate any network, regardless of the native connectivity.

Learning

To assess whether any learning took place over the course of these initial four training days, we examined the correlations between the two target regions (which had been trained to increase connectivity), the two target-control pairs (which were trained to decrease connectivity), as well as the composite difference measure. Figure 3 shows the results of this analysis for the ASD group. As can be seen, over the course of the four training days, correlations between the two target regions steadily increased (with a significant difference between most days, $P = 4x10^{-4}$ between day1 and day4, mean change in correlation = 0.11, Fig. 3A), while correlations between target1 and control decreased (significant difference between day1 and all other days, $P = 8x10^{-4}$ between day1 and day4 mean change = -0.13, Fig. 3B). Though there is an overall decrease between target2 and control, this does not reach significance and the mean change is small (Fig. 3C). This is in line with the lack of differentiation between the ASD group and the TD group in target2-control correlations (Fig 1C), suggesting neurotypical network connectivity is more resilient to change. Figure 3D shows the overall composite difference measure, taking into account all three correlation pairs, where there is a strong and consistent increase between day1 and day4 ($P < 2x10^{-4}$, mean change = 0.19). 15 of the 17 participants showed a positive change in this measure (14/17 had a positive change in target1-target2 correlations, as well as a negative

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

change in target1-control correlations). The TD control group on the other hand, showed no significant changes between days in any of the three pairwise combinations, or in the composite difference measure. Only 4/10 participants in this group showed a change in the trained direction in the composite measure, within the range of chance, and the magnitude of change was minimal relative to the change seen in the ASD participants. Figure 4A shows these data for all the individual ASD participants, while Figure 4B shows the individual TD participants. The full results for all individual participants, for all days, are displayed in fig. S4. To further ensure that the differences in results between the ASD and the TD groups were not due to lower statistical power in the TD group because of the smaller sample size, we re-calculated the change in correlations for each of the possible subsets of 10 participants from our ASD group, and found that the significant difference between day1 and day4 was maintained for all subsets in the target-target correlations, for 83% of possible subsets in the target1-control correlations, and for all possible subsets in the composite correlation difference measure (p-value < 0.007 for all subsets).

We next set out to test how long this learning would be maintained. To address this question, we called back the participants for a follow up study, in which they returned for another, shorter round of training. To get a good indication of the persistence of the training effect, participants were called back in a staggered manner, from as little as 5 weeks and up to 56 weeks from their original training. Our results indicate that the learning was mostly, though not fully, preserved, even after such an extended time period (Fig. 3A-D, follow up). In fact, although there was variation between subjects in the degree of retention, there was no correlation between the time that had elapsed and the rate of retention (see fig. S5). Since there were only two feedback runs in this follow up scan, we also compared them to just the first two feedback runs for the first four days, in order to account for any differences arising from the different number of runs. The results using just the first two runs for the first four days were not in any way different from the results using the full data (see fig. S6).

Whole brain analysis

So far we had only considered what happens in the regions that were trained. In order to get a more comprehensive picture of the effects of the training on the brain, we conducted a whole

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

brain analysis, which looked for changes during the training period (i.e. from day1 to day4). We calculated three maps, one for target1, one for target2, and one for the control region, with each map showing the change from day1 to day4 in the correlation of each voxel in the brain to the corresponding region. We then carried out a t-test across all participants for each of these three analyses, and the resulting maps for the ASD group are displayed in Figure 5. The changes were exactly as predicted by the training: the strongest positive change in correlation to target1 over the training period was in the somatosensory cortex (with a peak at target2), and the strongest negative change was in the control region. Changes to correlations with target2 were seen in the STS with a peak in target1, and changes in correlation to control were seen in bilateral STS (Fig. 5A). Since we were training a network of three nodes, rather than a simple connection between two regions, we next calculated the composite change: for each voxel, the change between day1 and day4 in its correlation to target1 minus its correlation to control (Fig. 5B), and the same change in its correlation to target2 minus control (Fig. 5C). This analysis yielded similar but far stronger results. The maps of the composite correlations were corrected at a very conservative cluster threshold determined by random permutation testing, in accordance with recent statistical recommendations for analyses utilizing cluster size (Eklund et al., 2016) (see Methods). These results support a causative role for the feedback itself, as the specific relationship that was trained between the two targets and the control came up in completely independent, whole-brain analysis. That is, using target1 relative to the control seed, the largest change in the whole brain was found in target2, even though that region was not pre-selected and the analyses did not constrain this to happen, and vice-versa using target2 and control. Note that we do not expect to find changes between day1 and day4 in either target1 or control in the target1-control map, as these regions did not change in relation to themselves. Rather, this analysis highlights all the other areas, outside of those two regions, which changed their correlation over the course of training in relation to target1 (increasing) and to control (decreasing), finding the peak of this change in target2. The same is true for the target2-control map, which shows an even greater effect focused on target1, consistent with the ROI analysis results showing a greater decoupling of target1 from control than target2 from control. Note that figure 5 shows results only for the ASD group, as, no significant peaks were identified in any of the target or control regions for the TD control group, and no voxels survived the cluster correction threshold.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transfer to resting state following training

The training-related changes we have demonstrated to this point were during the neurofeedback scans themselves. To be of any potential clinical value, these changes must also generalize beyond the training sessions, to the resting state scans, which reflect the baseline connectivity of the brain when not engaged in a specific task. Changes were overall smaller than those seen during the training, but significant changes were found between day1 and day4 (target1-target2 correlations mean change = 0.07, $P<0.038$, composite correlations measure mean change = 0.1, $P<2x10^{-4}$), and between day1 and the follow up (target1-target2 mean change = 0.09, $P<0.011$, composite correlations measure mean change = 0.11, $P<2x10^{-4}$). Change in rest was significantly correlated to change during the neurofeedback scans ($r = 0.42$, $P<0.04$, permutation test). Moreover, 14/15 participants who came in for the follow up showed an increase in the composite correlation measure (Fig. 6A). To assess whether the changes seen in the follow up could simply be a function of the elapsed time, we examined data from all participants in the previous study (used to define the training regions) which had at least two resting state scans from two different time points, and evaluated the change in connectivity seen between the sessions. 19 participants had two such data points, and the average time between the sessions was 13.2 months. The change however was not significant for any of the pairwise correlations, or the composite correlation measure (Fig. 6A). We next looked at the composite measure for the 10 participants from our study who had also participated in the previous resting state experiment (and in the follow up), and compared change from the previous experiment to the first rest sessions on day1, before any training, and in this subset also the change from day1 of the training to day4 and to the follow up (Fig. 6B). While there was no significant change from the previous experiment to day1 (mean time interval = 38.3 months, mean change = -0.14), there was significant change to day4 (mean change = 0.1, $P<0.025$) and to the follow up (mean change = 0.11, $P<0.006$). Taken together, these analyses provide strong evidence that the changes we observed were not a function of elapsed time but rather occurred as a direct result of our neurofeedback regime.

Behavioral relevance

Finally, we asked whether the changes we see as a result of the training are in any way correlated to behavior. To this end, we looked at changes in behavior as measured by the behavioral

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

questionnaires filled out by the parents prior to training, and two weeks after the end of the initial training set. These behavioral results included two statistical outliers who were removed from the analysis (see Methods). We compared the change in these behavioral questionnaires to the change in correlations during the resting state scans on the first and last days. There were two measures of behavior: the first, the Social Responsiveness Scale (SRS), has previously been found to correlate with functional connectivity in this network (see section on target selection), and therefore the change in this rating was expected to correlate with the change in the network. The second, the Behavior Rating Inventory of Executive Function (BRIEF), measures executive function rather than social abilities, and though patients with ASD show deficits on this measure (Gioia et al., 2000), it is expected to reflect prefrontal functioning, and we expected that changes on this measure would not correlate to changes in the social network being trained (Anderson et al., 2002, Anderson et al., 2005, Mahone et al., 2009). Indeed, there was a significant correlation between changes in the resting state network and the change (pre training minus post training, so that a positive change corresponds to a reduction in symptoms) in SRS (Figure 7), but no such correlation with the BRIEF (change in SRS vs. change in resting state correlations: $r = 0.56$, $p = 0.016$; change in BRIEF vs. change in resting state correlations: $r = 0.09$, $p = 0.39$). We further tested for the correlation between the change in SRS vs. change in the resting state correlations, after partialing out the contribution of the BRIEF. After removing the variance explained by the BRIEF, the correlation between the rest and SRS was even higher ($r = 0.6$, $p = 0.02$), indicating that change in resting state correlations was captured by the change in SRS scores, but not the BRIEF. Note that the SRS was chosen as our behavioral measure because it has consistently been shown to correlate with aberrations in network structure in ASD (Gotts et al., 2012, Wallace et al., 2012, Di Martino et al., 2009b, Anderson et al., 2011). Nevertheless, it was not designed to be used with such a short test-retest interval (3 weeks), and has only been validated for intervals of 6 months or more (Hus et al., 2013, Constantino et al., 2003, Bolte et al., 2008). We can therefore make no claims regarding the absolute values of the change in scores.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

The study of rt-fMRI covert neurofeedback – feedback given on pre-specified brain activity, but without providing participants with further information regarding the nature of the task or an explicit strategy through which to control the feedback – is still in its infancy (for a review see (Sitaram et al., 2017)). These results not only help solidify the existing evidence that reward-mediated learning through covert neurofeedback is possible (Shibata et al., 2011, Ramot et al., 2016, Amano et al., 2016), but also expand our knowledge in important ways. In this study, we have demonstrated that covert neurofeedback can be used to modulate correlations between distinct, physically distant networks (Figure 3), in the great majority of participants (15 of17, Figure 4, Figure S4). We have further shown that this modulation is possible even in cases of aberrant network structure, in clinical populations, and is sustainable for extended periods of time (some of our follow up sessions were a year after the original training, and we saw no evidence for an effect of time as a modulator of retention, Figure S5).

It could be argued that the changes in functional connectivity that we found were not a result of the feedback, but rather of the multiple fMRI visits, or were somehow precipitated by the nature of the puzzle task. Though imperfect because of the different population, the control using the TD group, which received feedback largely orthogonal to the network trained in the ASD group, provided further evidence for the necessity of the feedback itself in inducing these changes. The TD group, which went through the exact same protocol as the ASD group but received feedback on a different network, did not demonstrate the changes in connectivity seen in the ASD group (Figure 4, Figure S4). Moreover, the extraordinary specificity of the changes revealed by the whole brain analysis (Figure 5), peaking exactly at the small ROIs that were chosen for the training, would also seem to preclude an alternative explanation. The further significance of the whole-brain analysis is that this learning is spatially specific even when training disparate networks, meaning that it is possible to target specific regions of the brain. However, it is important to note in this regard that though the peaks were centered on the regions we were training, the changes spread to entire networks, as would be expected from the architecture of the brain, which is composed of large-scale networks of multiple brain regions, making it difficult to induce changes to just one region in isolation.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In a larger context, the failure to induce change in the direction of training in the control subjects, suggests that while it is possible to train networks that are fundamentally connected, it is much more difficult to train networks that are uncorrelated or weakly correlated in the typically developing brain. This conclusion is also bolstered by the failure of the training to induce change in the ASD group between target2 and control, regions which did not differ in their connectivity from that of the TDs. It is also possible that some networks may be more difficult to train than others, as has previously been suggested (Harmelech et al., 2015). Future studies will be needed to better understand the basic constraints in modifying these relationships.

The prohibitive cost of rt-fMRI, or fMRI scans in general, limits the number of training sessions in these paradigms. Moreover, the under-connectivity between the target regions chosen here explains only some of the behavioral deficits, and is clearly not the sole underlying cause of autism. This study was therefore designed as a proof of principle that aberrant connectivity can be addressed through neurofeedback, rather than as a clinical intervention, and whatever behavioral effects we found were expected to be modest. Once such a causal relationship is established, future potential clinical applications might pursue more cost effective options, such as identifying EEG signatures that correspond to activity from these areas, as several groups have already begun to develop (Zotev et al., 2014, Meir-Hasson et al., 2014).

From a clinical perspective, the most important result in this study is the successful transfer of the change in correlations to the resting state. By itself, change during training does not guarantee generalization of the learning, or in this case of the change in the network structure. Though the change in the resting state was somewhat more modest than the change seen during training, it was reliable and consistent, and could not be explained simply by the passage of time (Figure 6). The behavioral results, though preliminary in their scope and limited by the timescale on which they were measured, demonstrate that change in behaviorally relevant networks correlates with change in behavior, which is a crucial and entirely non-trivial point in terms of the potential clinical applications of neurofeedback. This finding also adds to the debate regarding the nature of the functional and structural hypoconnectivity found in ASD, whether it plays a causative role in ASD or is simply a downstream effect (Vasa et al., 2016). A change in behavior following a change in connectivity suggests the former, bolstering the mechanistic approach to functional connectivity in ASD. However, it should be noted that underconnectivity is not the full story in ASD, and there is growing evidence for hyper cortico-thalamic

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

connectivity, alongside the cortico-cortico hypoconnectivity (Picci et al., 2016b). Future studies will have to replicate and expand on the behavioral findings, but this method of testing behavioral changes following connectivity changes could be a promising tool for assessing different models of autism as well as other neuropsychiatric disorders.

The targets used in this paradigm were derived from a group analysis, and were not individually localized. As the variance within groups suggests, this is likely not the optimal method for ROI selection, especially if the focus is on behavioral change. These results are therefore probably an under estimation of what this tool can do, with individually tailored ROIs. It is not clear what the best method for individual ROI selection would be, or which localizer would best identify target regions, but it is another direction that should be pursued in future studies.

Since the lack of an explicit strategy allows covert neurofeedback to be used to directly target all manner of abstract, behaviorally relevant networks, potential applications could be far ranging, encompassing many clinical disorders with underlying aberrant connectivity at their core. Moreover, this is a promising technique to be used in more basic science questions, as a tool to investigate questions of causality.

Methods

Participants:

19 Males aged 15-25 (mean age = 20.93) who met the DSM-IV criteria for autistic disorder, an autism cut-off score for social symptoms on the Autism Diagnostic Review (ADR) and/or an ASD cutoff score from social+communication symptoms on the Autism Diagnostic Observation Schedule (ADOS), all administered by a trained, research-reliable clinician, were recruited for this experiment. Additionaly, 11 age matched typically developing males were recruited for the control group. All participants had normal to corrected to normal vision. IQ scores were obtained for all participants, and all full-scale IQ scores were ≥85 as measured by the Wechsler Abbreviated Scale of Intelligence, the Wechsler Adult Intelligence Scale-III, or the Wechsler Intelligence Scale for Children-IV. Participant groups did not differ in terms of full-scale IQ.

1 ASD participant was removed due to discomfort in scanner on day1, and another ASD participant was removed on day 2 due to anxiety. 1 TD participant was removed after day1 for excessive motion. 17 ASD and 10 TD participants completed all four days of neurofeedback

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

training. 15 ASD participants returned for the follow up experiment. The experiment was approved by the NIMH Institutional Review Board. Written informed consent was obtained from all participants.

Definition of ROIs:

3 Regions of Interest (ROIs) were selected for training: two targets and one control. The targets were chosen according to previous research as those with a large degree of reduced connectivity in Autism Spectrum Disorder compared with typically developing (TD) controls, based on between group analysis as explained in (Gotts et al., 2012). For this analysis, we used an expansion of the dataset published in (Gotts et al., 2012), N = 56 ASD, 62 TD. Of the 56 ASDs in this dataset, 11 participated in the neurofeedback study. Additional constraints placed on the choice of ROIs was for them to be physically distant from each other, and in different networks (see (Gotts et al., 2012) for details). All ROIs were defined as spheres of 4mm radius surrounding the focal points: Target1 - left Superior Temporal Sulcus (Talairach coordinates: -49, -29, 0), Target2 - left somatosensory cortex (Talairach coordinates: -54, 14, 39), and Control - right Inferior Parietal Lobe (chosen to be as uncorrelated with these two targets as possible in the TD dataset, Talairach coordinates: 49, -50, 42). Fig. 1 shows the between group difference in the correlations between the ROIs.

Imaging data collection and MRI parameters:

All scans were collected at the Functional Magnetic Resonance Imaging Core Facility on an 8 channel coil GE 3T (GE Signa HDxT 3.0T) magnet and receive-only head coil, with online slice time correction and motion correction. The scans included a 5 minute structural scan (MPRAGE) for anatomical co-registration, which had the following parameters: TE = 2.7, Flip Angle = 12, Bandwidth = 244.141, FOV = 30 (256 x 256), Slice Thickness = 1.2, axial slices. EPI was conducted with the following parameters: TR = 2s, Voxel size 3.2*3.2*3.2, Flip Angle: 60, TE = 30ms, Matrix = 72x72, Total TRs = 270, Slices: 37. All scans used an accelerated acquisition (GE's ASSET) with a factor of 2 in order to prevent gradient overheating.

Neurofeedback Experiment:

The initial neurofeedback experiment consisted of 4 training sessions over 8 days. There were 2 consecutive training days, a 6 day delay, then a final set of 2 consecutive training days. Each

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

training day had 2 initial rest scans, 4 neurofeedback sessions, and 2 final rest scans. All scans were 9 minutes long. Participants were instructed to maintain an eyes-open rest and look at the blank screen. Neuropsychological tests were administered at two timepoints: on the first training day before scanning, and two weeks following the last training day.

Follow-Up Experiment:

Follow up scans were conducted 5-56 weeks after the final training day and consisted of a single, abbreviated neurofeedback session with 2 rest scans followed by 2 neurofeedback sessions.

Online real-time data collection:

Regions of Interest (ROIs) were defined in Talairach space as described above. The standard Talairach brain was then co-registered to the structural scan collected that day, which was in turn co-registered to a short (10 TRs) functional echo-planar imaging scan (setup EPI) collected for that purpose each day before the first resting state session, to bring the ROIs into the native space during neurofeedback processing. All coregistration was carried out with the AFNI (Analysis of Functional Neuro-Images) software package (Cox, 1996).

Real-time fMRI algorithm:

During online processing of the data, 3D motion correction and slice time correction were carried out on all functional images. BOLD signal was extracted from each voxel in the ROIs and the mean signal was calculated for each ROI.

Feedback decisions were determined by a difference measure, taking into account both the changes in the trend between the two target ROIs and the control ROI. This difference measure was calculated for each TR and for each of the three ROIs. Our rt-fMRI algorithm calculated the difference between the mean signal in the current TR minus the signal in the preceding TR, giving the signal trend in each ROI (increasing or decreasing). If the trend in the two targets was the same, and opposite from the trend in the control ROI, then feedback was given, meaning both conditions had to be fulfilled for feedback to be given:

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\frac{ms(Target1(TR=t))-ms(Target1(TR=t-1))}{ms(Target2(TR=t))-ms(Target2(TR=t-1))} > 0 \quad \& \quad \frac{ms(Target1(TR=t))-ms(Target1(TR=t-1))}{ms(Control(TR=t))-ms(Control(TR=t-1))} < 0$$

(ms = mean signal)

Neurofeedback Procedure:

Each training session had 4 neurofeedback training scans. The scans started out with a uniformly grey screen. Participants were told that there is a picture hidden underneath, and were instructed to try to unveil the image during what was described as the puzzle task. Importantly, no further cognitive strategies or suggestions were given to the participant for the duration of the experiment. Participants were not informed that their performance on the puzzle task was determined by brain activation.

Neurofeedback Stimuli:

Participants received two forms of positive reinforcement whenever the real-time algorithm determined that the network requirements had been met: a "puzzle piece", i.e. a square of the hidden picture, would become visible on the screen with a concomitant sound of positive valence. This feedback was chosen to maximize engagement with the paradigm during the scan by providing a complex and interesting visual stimulus in a game-like setting, and the auditory stimulus was paired to ensure that participants would be aware of positive feedback independent of their visual attendance.

Visual Stimuli:

During rest scans participants were shown a uniformly grey screen.

During Neurofeedback training participants would begin with a uniformly grey screen. Then the image would become visible in small rectangular blocks, described to the participants as "puzzle pieces." There were 25 "puzzle pieces" per board, which would be displayed piece by piece until a whole image was unveiled. After a full board was completed, the screen would become blank and a new image would begin to appear. The images were randomly selected from a set of 100 non-social images devoid of people or text, like a landscape or an abstract painting.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

At the end of each 9 minute training round, participants viewed a scoreboard which told them how many individual pieces they had unveiled that round, as well as the top score that they had received that day. Participants were financially incentivized to beat their best score for that day.

fMRI offline data preprocessing:

Post-hoc signal preprocessing was conducted in AFNI. The first four EPI volumes from each run were removed to ensure remaining volumes were at magnetization steady state, and remaining large transients were removed through a squashing function (AFNI's 3dDespike). Volumes were slice-time corrected and motion parameters were estimated with rigid body transformations. Volumes were coregistered to the anatomical scan. Volumes were smoothed with 6mm blurring and normalized by the mean signal intensity of each voxel. The AFNI ANATICOR procedure was then applied to remove nuisance physiological and nonphysiological artifacts from the data (Jo et al., 2010). The anatomical scan was segmented into tissue compartments with Freesurfer (Fischl et al., 2002), Ventricle and white-matter masks were created and applied to the volume-registered EPI. Prior to smoothing, these masks gave pure nuisance times series for the ventricles and local estimates of the BOLD signal in white matter, averaged within a 15-mm radius. The measured respiration and heart rate signals were used to create Retroicor (Glover et al., 2000) and respiration volume per time (RVT) regressors (Birn et al., 2008). All nuisance time series in every run (average ventricle time series, average local white matter time series, 6 parameter estimates for head motion, and thirteen RVT and Retroicor regressors) were detrended with fourth-order polynomials before least-squares model fit to each voxel time series. No other filtering of the data was done. All participant data was aligned by affine registration to AFNI's TT-N27 template in standardized Talairach and Tournoux (Talairach and Tournoux, 1988) space.

Neuropsychological tests:

Baseline neuropsychological tests were conducted before the initial training session, and post-experiment surveys were collected two weeks after the final neurofeedback session. Parents filled out the Social Behavior Scale (SRS) to identify common social behaviors in autism, as well as the Behavioral Rating Inventory of Executive Function (BRIEF). The 'informant' report (filled in by a parent) was used as it has been shown to be more accurate (McMahon and

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Solomon, 2015). An independent dataset of ASD subjects who did not participate in this experiment but had SRS test-retest data was used to determine change reliability. Data points that were beyond 3 standard deviations from the mean as determined by this analysis were excluded.

Cognitive Strategy Questionnaire

We developed a cognitive strategy questionnaire that was completed by 11 of the 17 participants. Following their final scan session on day4 of the training, each of these participants was asked what they thought the experiment was about. Participants were then asked what they were doing during the scans, and if they used a particular cognitive strategy.

Finally, participants were asked to rate on a scale of 1-10 how hard they had been trying to solve the puzzles each day, how satisfied they felt when a new puzzle piece came up, and if there were differences between days. The first six participants did not complete this questionnaire, but were interviewed after the final scan and reported no knowledge of the objective of the task, and similar cognitive strategies to those later reported in the questionnaire. See Table S1 for the data from these questionnaires.

Data Analysis:

All data were analyzed with in-house software written in Matlab, as well as the AFNI software package. Data on the cortical surface were visualized with SUMA (SUrface MApping) (Saad et al., 2004). The composite difference measure was computed by subtracting the average correlation of the two target/control pairs, from the target/target correlation:

$$corr(Target1, Target2) - \frac{1}{2}(corr(Target1, Control) + corr(Target2, Control))$$

All p-values for the changes in correlation between days were computed through permutation tests, randomly permuting the days for 5000 iterations.

Whole-brain analysis:

For each participant, for each neurofeedback scan on day1 and day4, we first transformed the correlation values with Fisher's z-transform to improve normality, then calculated a difference measure per voxel: corr(voxel time series, avg. Target1 time series) - corr(voxel time series, avg. Control time series). The resulting maps held information regarding each voxel's differential

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlation to the Target1 vs. Control ROIs. We then averaged the maps for each participant across all 4 neurofeedback scans for each of the two days, and subtracted the average day1 map from the average day4 map. Each voxel in the resulting map now signified the change in correlation from day1 to day4, in the differential correlation to the Target1 ROI vs. the Control ROI, where a positive value means that this voxel was differentially more correlated to Target1 than to Control on day4 compared with day1. Normality of these data were ascertained using Lilliefor's goodness of fit test. We then carried out a t-test across the 17 participants, to identify voxels with a consistent change across subjects. Maps were corrected using a permutation test to determine significant cluster size, with day1 and day4 randomly permuted for each participant across 5000 permutations (as suggested by (Eklund et al., 2016)). These permutations were carried out at p-value thresholds of 0.05, 0.01, 0.005, 0.001 and 0.0005, and a mask was created of voxels that survived any of these corrections. The mask was then applied to the map shown in Fig. 5A, which was set at a p-value threshold of 0.05.

The same procedure was carried out for the Target2 minus Control differential correlation, and the resulting map is shown in Fig. 5B.

Data availability

Data will be made available in a public data repository such as XNAT.

Acknowledgments: We thank Lauren Kenworthy for insights into behavioral testing methods in ASD, and Miriam Menken for help with data pre-processing. This work was supported by the Revson Foundation Women in Science award through the Weizmann Institute of Science (to M.R) and by the Intramural Research Program, National Institute of Mental Health (ZIAMH002920).

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References and Notes:

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

ALLISON, T., PUCE, A. & MCCARTHY, G. 2000. Social perception from visual cues: role of the STS region. *Trends Cogn Sci,* 4**,** 267-278.

AMANO, K., SHIBATA, K., KAWATO, M., SASAKI, Y. & WATANABE, T. 2016. Learning to Associate Orientation with Color in Early Visual Areas by Associative Decoded fMRI Neurofeedback. *Curr Biol*.

ANDERSON, J. S., NIELSEN, J. A., FROEHLICH, A. L., DUBRAY, M. B., DRUZGAL, T. J., CARIELLO, A. N., COOPERRIDER, J. R., ZIELINSKI, B. A., RAVICHANDRAN, C. & FLETCHER, P. T. 2011. Functional connectivity magnetic resonance imaging classification of autism. *Brain,* 134**,** 3742-3754.

ANDERSON, V., JACOBS, R. & HARVEY, A. S. 2005. Prefrontal lesions and attentional skills in childhood. *Journal of the International Neuropsychological Society,* 11**,** 817-831.

ANDERSON, V. A., ANDERSON, P., NORTHAM, E., JACOBS, R. & MIKIEWICZ, O. 2002. Relationships between cognitive and behavioral measures of executive function in children with brain disease. *Child Neuropsychology,* 8**,** 231-240.

BELMONTE, M. K., ALLEN, G., BECKEL-MITCHENER, A., BOULANGER, L. M., CARPER, R. A. & WEBB, S. J. 2004. Autism and abnormal development of brain connectivity. *J Neurosci,* 24**,** 9228-31.

BIRBAUMER, N., RUIZ, S. & SITARAM, R. 2013. Learned regulation of brain metabolism. *Trends Cogn Sci,* 17**,** 295-302.

BIRN, R. M., SMITH, M. A., JONES, T. B. & BANDETTINI, P. A. 2008. The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *Neuroimage,* 40**,** 644-54.

BOLTE, S., POUSTKA, F. & CONSTANTINO, J. N. 2008. Assessing autistic traits: cross-cultural validation of the social responsiveness scale (SRS). *Autism Res,* 1**,** 354-63.

CHEN, C. P., KEOWN, C. L., JAHEDI, A., NAIR, A., PFLIEGER, M. E., BAILEY, B. A. & MULLER, R. A. 2015. Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism. *Neuroimage Clin,* 8**,** 238-45.

CHENG, W., ROLLS, E. T., GU, H., ZHANG, J. & FENG, J. 2015. Autism: reduced connectivity between cortical areas involved in face expression, theory of mind, and the sense of self. *Brain***,** awv051.

CONSTANTINO, J. N., DAVIS, S. A., TODD, R. D., SCHINDLER, M. K., GROSS, M. M., BROPHY, S. L., METZGER, L. M., SHOUSHTARI, C. S., SPLINTER, R. & REICH, W. 2003. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *J Autism Dev Disord,* 33**,** 427-33.

CORTESE, A., AMANO, K., KOIZUMI, A., LAU, H. & KAWATO, M. 2017. Decoded fMRI neurofeedback can induce bidirectional confidence changes within single participants. *Neuroimage,* 149**,** 323-337.

COX, R. W. 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res,* 29**,** 162-73.

DAMASIO, A. R., GRABOWSKI, T. J., BECHARA, A., DAMASIO, H., PONTO, L. L. B., PARVIZI, J. & HICHWA, R. D. 2000. Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience,* 3**,** 1049-1056.

DEBETTENCOURT, M. T., COHEN, J. D., LEE, R. F., NORMAN, K. A. & TURK-BROWNE, N. B. 2015. Closed-loop training of attention with real-time brain imaging. *Nat Neurosci,* 18**,** 470-5.

DI MARTINO, A., ROSS, K., UDDIN, L. Q., SKLAR, A. B., CASTELLANOS, F. X. & MILHAM, M. P. 2009a. Functional brain correlates of social and nonsocial processes in autism spectrum disorders: an activation likelihood estimation meta-analysis. *Biological psychiatry,* 65**,** 63-74.

DI MARTINO, A., SHEHZAD, Z., KELLY, C., ROY, A. K., GEE, D. G., UDDIN, L. Q., GOTIMER, K., KLEIN, D. F., CASTELLANOS, F. X. & MILHAM, M. P. 2009b. Relationship between cingulo-insular functional connectivity and autistic traits in neurotypical adults. *American Journal of Psychiatry,* 166**,** 891-899.

DI MARTINO, A., YAN, C. G., LI, Q., DENIO, E., CASTELLANOS, F. X., ALAERTS, K., ANDERSON, J. S., ASSAF, M., BOOKHEIMER, S. Y., DAPRETTO, M., DEEN, B., DELMONTE, S., DINSTEIN, I., ERTL-WAGNER, B., FAIR, D. A., GALLAGHER, L., KENNEDY, D. P., KEOWN, C. L., KEYSERS, C., LAINHART, J. E., LORD, C., LUNA, B., MENON, V., MINSHEW, N. J., MONK, C. S., MUELLER, S., MULLER, R. A., NEBEL, M. B., NIGG, J. T., O'HEARN, K., PELPHREY, K. A., PELTIER, S. J., RUDIE, J. D., SUNAERT, S., THIOUX, M., TYSZKA, J. M., UDDIN, L. Q., VERHOEVEN, J. S., WENDEROTH, N., WIGGINS, J. L., MOSTOFSKY, S. H. & MILHAM, M. P. 2014. The autism brain

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry,* 19**,** 659-667.

EKLUND, A., NICHOLS, T. E. & KNUTSSON, H. 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A,* 113**,** 7900-5.

FISCHL, B., SALAT, D. H., BUSA, E., ALBERT, M., DIETERICH, M., HASELGROVE, C., VAN DER KOUWE, A., KILLIANY, R., KENNEDY, D., KLAVENESS, S., MONTILLO, A., MAKRIS, N., ROSEN, B. & DALE, A. M. 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron,* 33**,** 341-55.

FRITH, U. & FRITH, C. 2010. The social brain: allowing humans to boldly go where no other species has been. *Philos Trans R Soc Lond B Biol Sci,* 365**,** 165-76.

GIOIA, G. A., ISQUITH, P. K., GUY, S. C. & KENWORTHY, L. 2000. Behavior rating inventory of executive function. *Child Neuropsychol,* 6**,** 235-8.

GLOVER, G. H., LI, T. Q. & RESS, D. 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn Reson Med,* 44**,** 162-7.

GOTTS, S. J., SIMMONS, W. K., MILBURY, L. A., WALLACE, G. L., COX, R. W. & MARTIN, A. 2012. Fractionation of social brain circuits in autism spectrum disorders. *Brain,* 135**,** 2711-25.

HAHAMY, A., BEHRMANN, M. & MALACH, R. 2015. The idiosyncratic brain: distortion of spontaneous connectivity patterns in autism spectrum disorder. *Nat Neurosci,* 18**,** 302-9.

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**,** 2588-95.

HUS, V., BISHOP, S., GOTHAM, K., HUERTA, M. & LORD, C. 2013. Factors influencing scores on the social responsiveness scale. *J Child Psychol Psychiatry,* 54**,** 216-24.

JO, H. J., SAAD, Z. S., SIMMONS, W. K., MILBURY, L. A. & COX, R. W. 2010. Mapping sources of correlation in resting state FMRI, with artifact detection and removal. *Neuroimage,* 52**,** 571-82.

KANA, R. K., KELLER, T. A., CHERKASSKY, V. L., MINSHEW, N. J. & JUST, M. A. 2009. Atypical frontal-posterior synchronization of Theory of Mind regions in autism during mental state attribution. *Social neuroscience,* 4**,** 135-152.

KEOWN, C. L., DATKO, M. C., CHEN, C. P., MAXIMO, J. O., JAHEDI, A. & MÜLLER, R.-A. 2016. Network Organization Is Globally Atypical in Autism: A Graph Theory Study of Intrinsic Functional Connectivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

KHAN, S., GRAMFORT, A., SHETTY, N. R., KITZBICHLER, M. G., GANESAN, S., MORAN, J. M., LEE, S. M., GABRIELI, J. D. E., TAGER-FLUSBERG, H. B., JOSEPH, R. M., HERBERT, M. R., HAMALAINEN, M. S. & KENET, T. 2013. Local and long-range functional connectivity is reduced in concert in autism spectrum disorders. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 3107-3112.

KHAN, S., MICHMIZOS, K., TOMMERDAHL, M., GANESAN, S., KITZBICHLER, M. G., ZETINO, M., GAREL, K.-L. A., HERBERT, M. R., HÄMÄLÄINEN, M. S. & KENET, T. 2015. Somatosensory cortex functional connectivity abnormalities in autism show opposite trends, depending on direction and spatial scale. *Brain***,** awv043.

KOIZUMI, A., AMANO, K., CORTESE, A., SHIBATA, K., YOSHIDA, W., SEYMOUR, B., KAWATO, M. & LAU, H. 2016. Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour,* 1**,** 0006.

MAHONE, E. M., MARTIN, R., KATES, W. R., HAY, T. & HORSKA, A. 2009. Neuroimaging correlates of parent ratings of working memory in typically developing children. *Journal of the International Neuropsychological Society,* 15**,** 31-41.

MCMAHON, C. M. & SOLOMON, M. 2015. Brief report: parent-adolescent informant discrepancies of social skill importance and social skill engagement for higher-functioning adolescents with autism spectrum disorder. *J Autism Dev Disord,* 45**,** 3396-403.

MEIR-HASSON, Y., KINREICH, S., PODLIPSKY, I., HENDLER, T. & INTRATOR, N. 2014. An EEG Finger-Print of fMRI deep regional activation. *Neuroimage,* 102 Pt 1**,** 128-41.

MÜLLER, R.-A., PIERCE, K., AMBROSE, J. B., ALLEN, G. & COURCHESNE, E. 2001. Atypical patterns of cerebral motor activation in autism: a functional magnetic resonance study. *Biological psychiatry,* 49**,** 665-676.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MULLER, R. A., SHIH, P., KEEHN, B., DEYOE, J. R., LEYDEN, K. M. & SHUKLA, D. K. 2011. Underconnected, but How? A Survey of Functional Connectivity MRI Studies in Autism Spectrum Disorders. *Cerebral Cortex,* 21**,** 2233-2243.

OTERO, T. L., SCHATZ, R. B., MERRILL, A. C. & BELLINI, S. 2015. Social skills training for youth with autism spectrum disorders: a follow-up. *Child Adolesc Psychiatr Clin N Am,* 24**,** 99-115.

PICCI, G., GOTTS, S. J. & SCHERF, K. S. 2016a. A theoretical rut: revisiting and critically evaluating the generalized under/over-connectivity hypothesis of autism. *Developmental Science,* 19**,** 524-549.

PICCI, G., GOTTS, S. J. & SCHERF, K. S. 2016b. A theoretical rut: revisiting and critically evaluating the generalized under/over-connectivity hypothesis of autism. *Dev Sci,* 19**,** 524-49.

PLITT, M., BARNES, K. A., WALLACE, G. L., KENWORTHY, L. & MARTIN, A. 2015. Resting-state functional connectivity predicts longitudinal change in autistic traits and adaptive functioning in autism. *Proc Natl Acad Sci U S A,* 112**,** E6699-706.

RAMOT, M., GROSSMAN, S., FRIEDMAN, D. & MALACH, R. 2016. Covert neurofeedback without awareness shapes cortical network spontaneous connectivity. *Proc Natl Acad Sci U S A,* 113**,** E2413-20.

SAAD, Z. S., REYNOLDS, R. C., ARGALL, B., JAPEE, S. & COX, R. W. 2004. Suma: An interface for surface-based intra- and inter-subject analysis with AFNI. *2004 2ND IEEE INTERNATIONAL SYMPOSIUM ON BIOMEDICAL IMAGING: MACRO TO NANO, VOLS 1 and 2***,** 1510-1513.

SHIBATA, K., WATANABE, T., SASAKI, Y. & KAWATO, M. 2011. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science,* 334**,** 1413-5.

SITARAM, R., ROS, T., STOECKEL, L., HALLER, S., SCHARNOWSKI, F., LEWIS-PEACOCK, J., WEISKOPF, N., BLEFARI, M. L., RANA, M., OBLAK, E., BIRBAUMER, N. & SULZER, J. 2017. Closed-loop brain training: the science of neurofeedback. *Nat Rev Neurosci,* 18**,** 86-100.

STOECKEL, L. E., GARRISON, K. A., GHOSH, S., WIGHTON, P., HANLON, C. A., GILMAN, J. M., GREER, S., TURK-BROWNE, N. B., DEBETTENCOURT, M. T., SCHEINOST, D., CRADDOCK, C., THOMPSON, T., CALDERON, V., BAUER, C. C., GEORGE, M., BREITER, H. C., WHITFIELD-GABRIELI, S., GABRIELI, J. D., LACONTE, S. M., HIRSHBERG, L., BREWER, J. A., HAMPSON, M., VAN DER KOUWE, A., MACKEY, S. & EVINS, A. E. 2014. Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *Neuroimage Clin,* 5**,** 245-55.

SULZER, J., HALLER, S., SCHARNOWSKI, F., WEISKOPF, N., BIRBAUMER, N., BLEFARI, M. L., BRUEHL, A. B., COHEN, L. G., DECHARMS, R. C., GASSERT, R., GOEBEL, R., HERWIG, U., LACONTE, S., LINDEN, D., LUFT, A., SEIFRITZ, E. & SITARAM, R. 2013. Real-time fMRI neurofeedback: progress and challenges. *Neuroimage,* 76**,** 386-99.

TALAIRACH, J. & TOURNOUX, P. 1988. *Co-planar stereotaxic atlas of the human brain : 3-dimensional proportional system : an approach to cerebral imaging,* Stuttgart ; New York, Georg Thieme.

TUTTLE, A. H., BARTSCH, V. B. & ZYLKA, M. J. 2016. The Troubled Touch of Autism. *Cell,* 166**,** 273-4.

VASA, R. A., MOSTOFSKY, S. H. & EWEN, J. B. 2016. The Disrupted Connectivity Hypothesis of Autism Spectrum Disorders: Time for the Next Phase in Research. *Biol Psychiatry Cogn Neurosci Neuroimaging,* 1**,** 245-252.

WALLACE, G. L., SHAW, P., LEE, N. R., CLASEN, L. S., RAZNAHAN, A., LENROOT, R. K., MARTIN, A. & GIEDD, J. N. 2012. Distinct cortical correlates of autistic versus antisocial traits in a longitudinal sample of typically developing youth. *J Neurosci,* 32**,** 4856-60.

WEISKOPF, N. 2012. Real-time fMRI and its application to neurofeedback. *Neuroimage,* 62**,** 682-92.

WILLIAMS WHITE, S., KEONIG, K. & SCAHILL, L. 2007. Social skills development in children with autism spectrum disorders: a review of the intervention research. *J Autism Dev Disord,* 37**,** 1858-68.

ZOTEV, V., PHILLIPS, R., YUAN, H., MISAKI, M. & BODURKA, J. 2014. Self-regulation of human brain activity using simultaneous real-time fMRI and EEG neurofeedback. *Neuroimage,* 85 Pt 3**,** 985-95.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

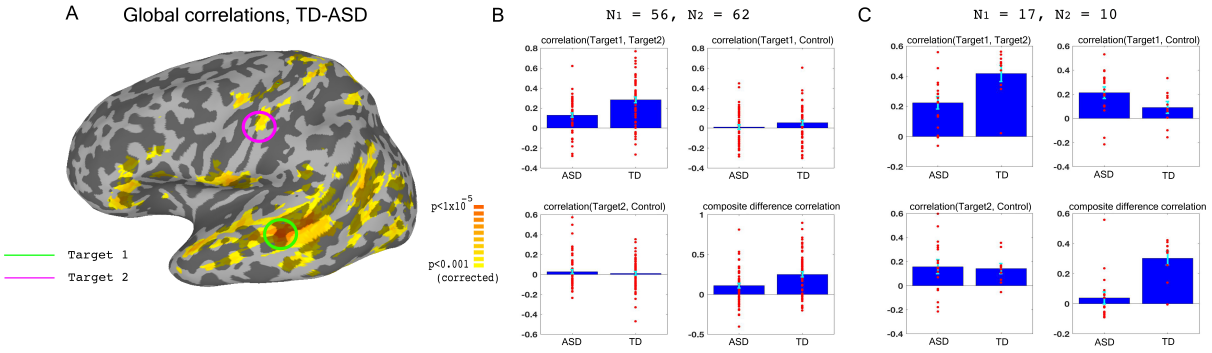

Fig. 1. Choosing ROIs. (A) Groups differences between TD (N = 62) and ASD participants (N = 56), matched for motion, age and IQ. Difference maps calculated on the average correlation of each voxel with all other grey matter voxels in the brain. Target1 was chosen as the region with the greatest between-group difference, and Target2 was chosen as the region in which the difference in connectivity to Target1 was greatest between groups, while also being in a physically distant, distinct network based on(Gotts et al., 2012). (B) Pairwise correlations for the dataset shown in (A), between the two targets (top left), target1 and control (top right), target2 and control (bottom left), and the composite difference measure, based on the difference in correlations between the two targets and the target-control pairs (see Methods). Difference between the ASD and TD groups is significant for target1-target2 correlations, and the composite difference measure. (C) Same as (B), but for the current cohort of participants who took part in the neurofeedback study. Correlations are averaged across the first two rest scans of the first day, before training. Between group difference is significant for target1-target2 correlations, target1-control correlations, and the composite difference measure. Blue bars represent the subject mean, cyan error bars mark +/- SEM. Red dots represent each individual subject.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

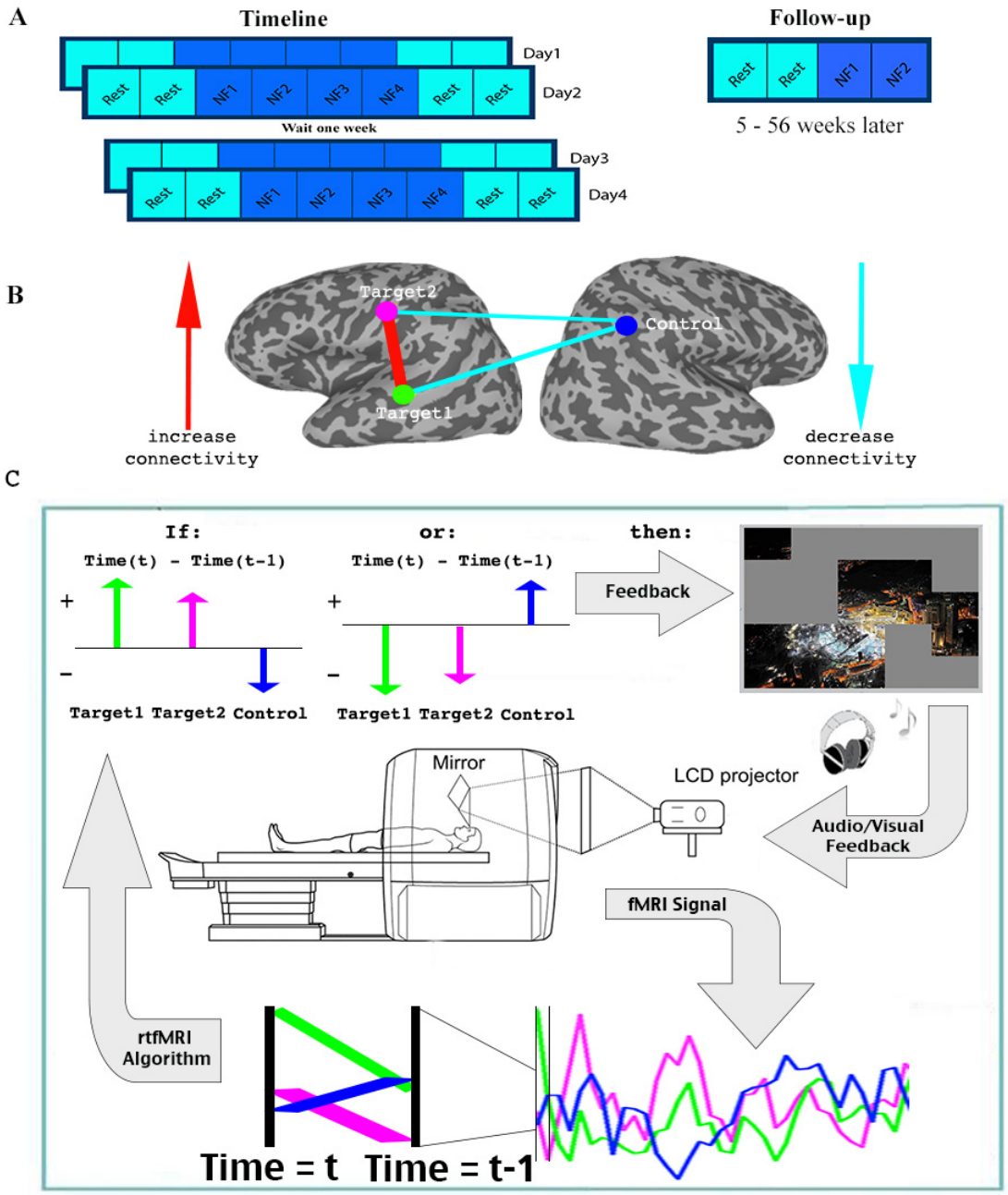

Fig. 2. Experimental paradigm. A. Timeline. B. Location of ROIs, and network being trained. C. Feedback session. Data was collected and analyzed in real-time, and a decision whether to present feedback (reveal a square of the picture + positive sound) was made based on the change in signal from one time (t-1) to the next (t) in the three ROIs. Feedback was given if the direction of change in the two targets was the same, and opposite from the direction of change in the control ROI.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

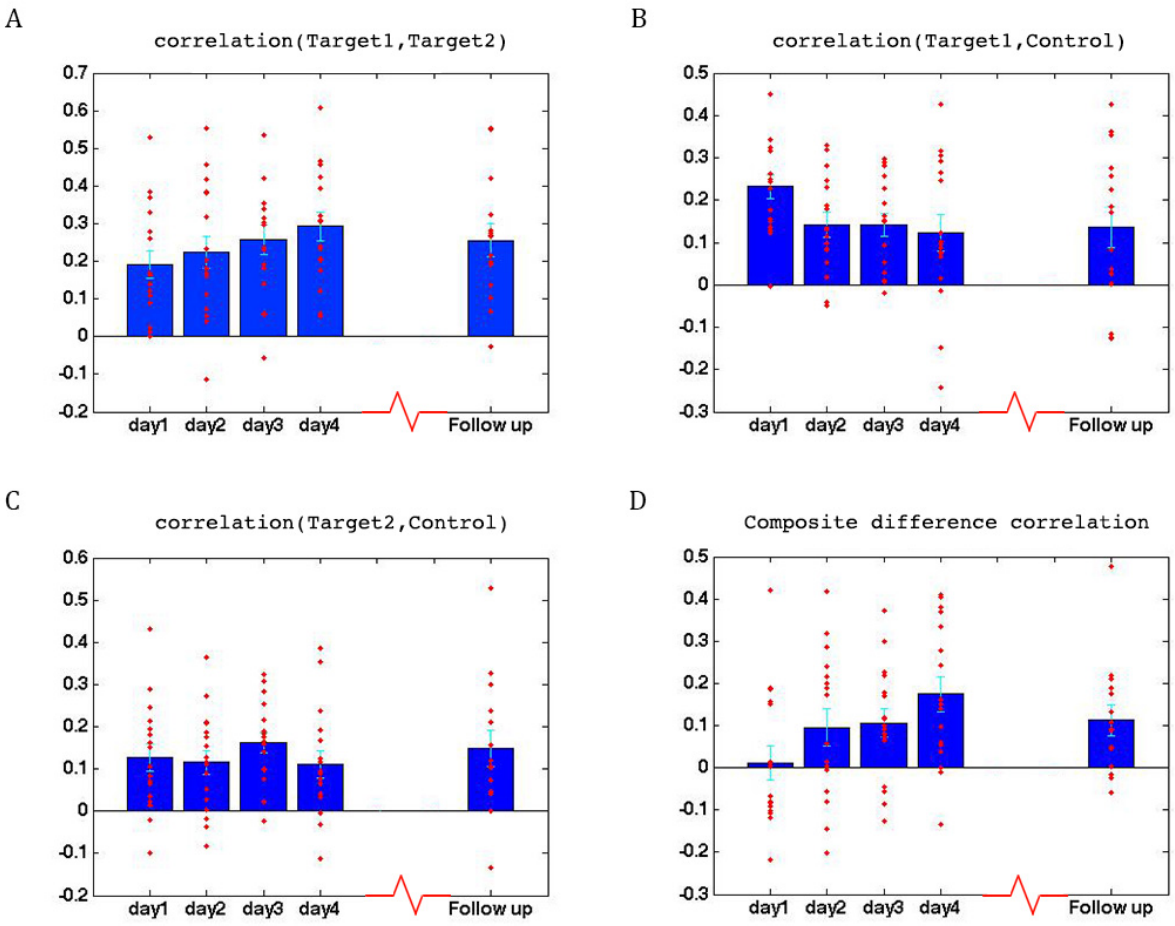

Fig. 3. Learning across days, ASD group. (A) Correlations between the two target regions per day, averaged across all four neurofeedback scans per day. Blue bars represent the subject mean, cyan error bars mark +/- SEM. Red dots represent each individual subject. The difference in correlations between day1 and all other days is significant, as is the difference between day2 and day4. (B) Correlations between Target1 and Control. There is a significant change between day1 and all other days. (C) Correlations between Target2 and Control. (D) Composite difference measure, showing the difference between target-target and target-control correlation pairs (see Methods). Day1 correlations are significantly different from all other days, and day2 is significant different from day4. In all panels N=17 for days 1-4, N=15 for the follow up. All p-values for differences between days were determined by permutation tests.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

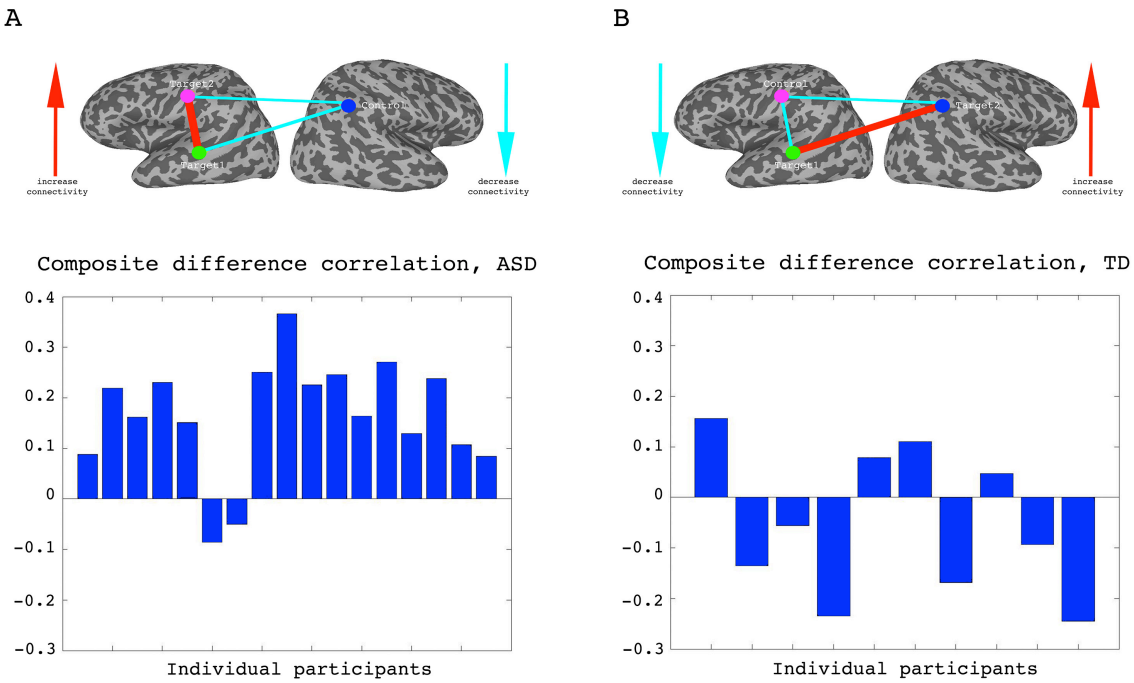

Fig. 4 Individual participant data. (A) Difference in the composite difference correlation from day1 to day4, for each of the 17 individual ASD participants averaged across all four neurofeedback scans per day, presented chronologically in order of scanning. (B) Same analysis for each of the 10 TD participants, presented chronologically. Note that the composite difference measure is comprised of target-target correlations minus the target-control pairs, and that the definition of targets and control differed between the two groups.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

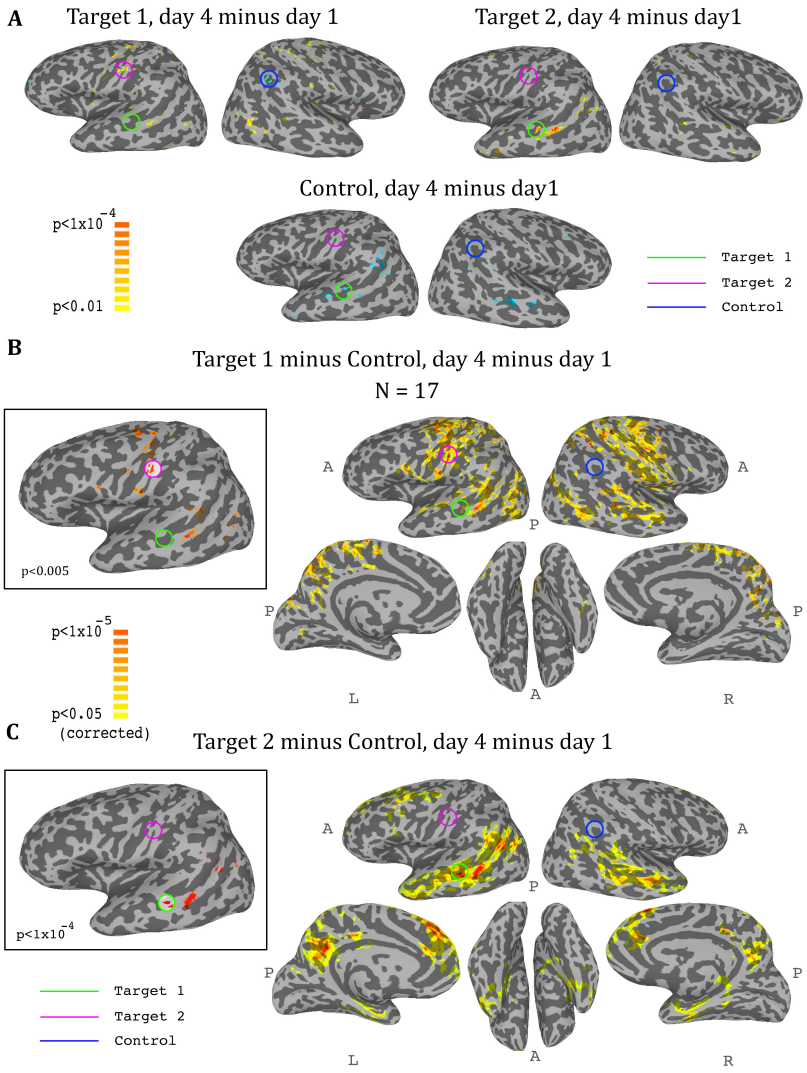

Fig. 5. Whole-brain analysis during neurofeedback, ASD group. (A) Top left: change in correlations to Target1, between day1 and day4, t-test across participants. High values represent voxels that showed a consistent change between day1 and day4, such that on day4 they were more correlated to Target1 than they were on day1. Note the positive peak in target2, and the negative peak in the control region. Top right: change in correlations to Target2, between day1 and day4, t-test across participants. Note the positive peak in target1. Bottom: change in correlations to the control region, between day1 and day4, t-test across participants. Note the negative peak in target1 and bilateral STS. (B) Change in differential correlation to the Target1 and Control ROIs, between day1 and day4, t-test across participants. High values represent voxels that showed a consistent change between day1 and day4, such that on day4 they were more correlated to Target1 and less correlated to Control than they were on day1. Inset shows the same analysis at a higher threshold. (C) Same as (B), for Target2 and Control. Note that for both maps, the other target, which was not included in the analysis, emerges as the area of greatest change across training days. Maps corrected through permutation tests (see Methods).

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

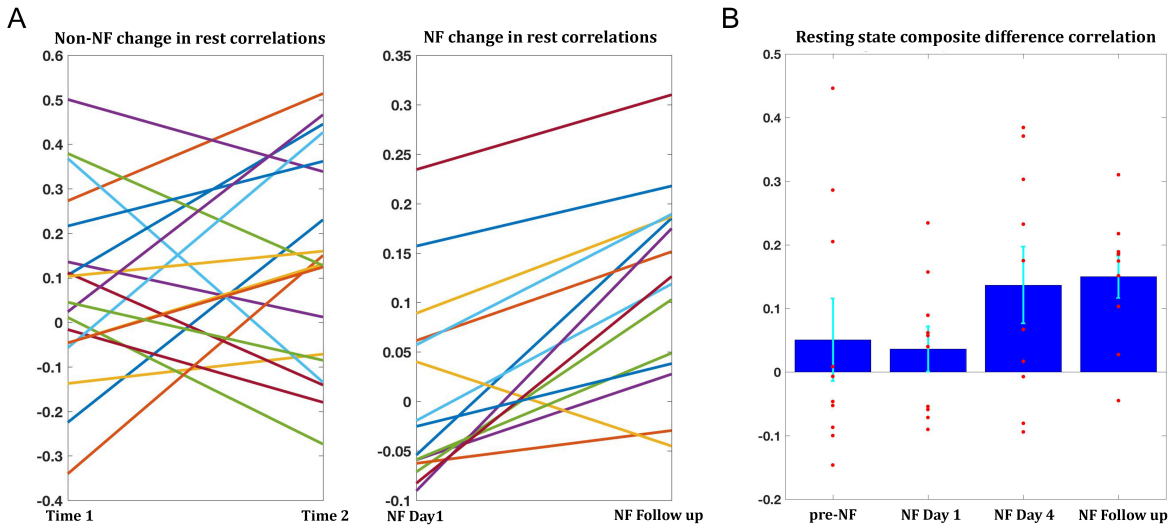

Fig. 6. Changes in resting state correlations. (A) Left panel shows the changes in resting state composite difference correlations for the 19 participants for which two previous data points were available from a previous study, prior to neurofeedback (average time between sessions 13.2 months). The right panel shows the change in resting state composite difference correlations from the very first pre-training rest sessions on the first day of neurofeedback, to the rest sessions collected in the follow up session (also before the neurofeedback training sessions that day), for the 15 participants who took part in the follow up session. Average time between sessions for this group was 6.2 months. (B) Changes in resting state composite difference correlations for the 10 participants who had data from both the previous study, and the follow up session. Change between neurofeedback day1 and neurofeedback day4 as well as neurofeedback follow up, are significant. All correlations are taken from the resting state scans at the beginning of the relevant session. NF = neurofeedback.

bioRxiv preprint doi: https://doi.org/10.1101/139824; this version posted May 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

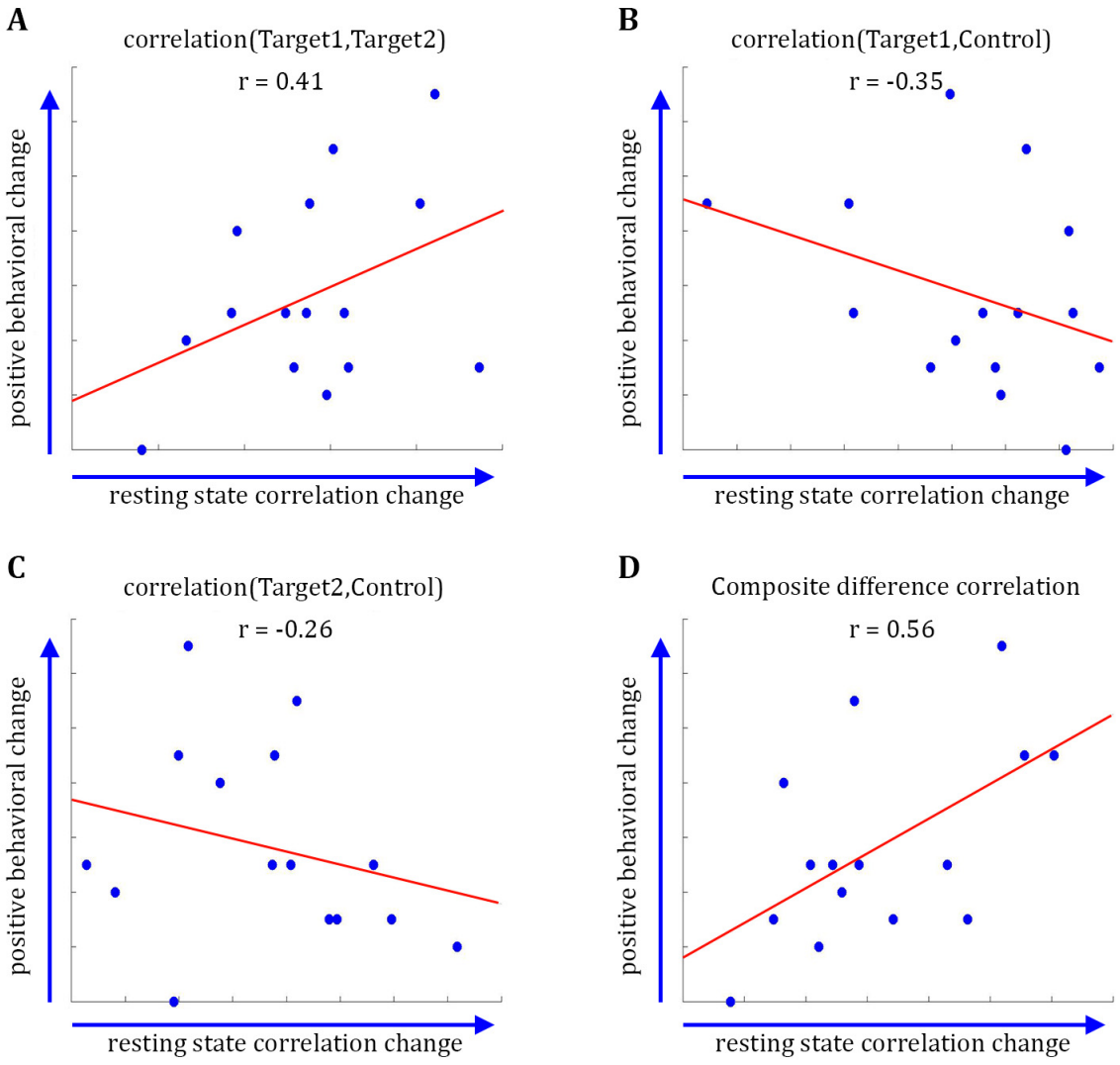

Fig. 7. Behavioral changes. Correlation between the change in the behavioral measure (SRS) score before and after training, and the change in resting state connectivity from the post-training rest scans on day4 to the rest scans on day1. (A) Behavioral change vs. change in Target1-Target2 correlations. (B) Behavior vs. Target1-Control. (C) Behavior vs. Target2-Control. (D) Behavior vs. composite difference corr. R values represent Pearson's correlation.