Supplementary Materials

Method

Participant Information

Patients were recruited through local advertisement and clinical referral from local psychiatric and psychological services. OCD diagnosis was confirmed by the referring clinician, or where recruitment was conducted through advertisement, by a consultant psychiatrist. Patients who met the MINI International Neuropsychiatric Interview (MINI, 36) criteria for OCD and scored 12 or more on the Yale-Brown Obsessive-Compulsive Scale (YBOCS, 37) were included in the study. We additionally imposed a cut-off of maximum score of 16 on the Montgomery-Åsberg Depression Rating Scale (MADRS, 38) during screening. Nonetheless, three patients reported MADRS scores in excess of this cut-off on the day of testing (scoring 20, 25 and 25, respectively). Omitting these patients from the analyses does not alter the results, and we therefore include these patients in the main analyses. Anxiety levels were quantified using the State-Trait Anxiety Inventory (STAI, 39), and in addition to the YBOCS, we collected a self-report scale of OCD symptomatology using the Obsessive-Compulsive Inventory - Revised (OCI-R, 40). Fourteen patients were un-medicated. Twenty-one of the 23 medicated patients were taking selective serotonin reuptake inhibitors (SSRIs); 14 were taking SSRIs in isolation, and 7 receiving a combination treatment, that in some cases included tricyclic antidepressants (TCAs) and neuroleptics (D2 antagonists). The two remaining patients were not taking SSRIs were taking TCAs in addition to neuroleptics. This study was approved by the Cambridge Central Research Ethics Committee.

Procedure

Participants completed two runs of a practice task prior to the experiment. These practice runs did not use shocks, but rather an image of a lightning bolt that subjects were told symbolized the shock they would receive in the real experiment. One practice was completed outside the scanner, where keyboard presses could avoid the presentation of the lightning bolt. A second practice was conducted while subjects were lying down in the scanner, using the foot-box between their feet to avoid lightning bolts. If participants were missed more than 25% of avoidance trials (i.e. did not make a response on time) during either practice session, the CS presentation time was increased by 100ms from the baseline of 750ms and this step was repeated until they reached this criterion. There were no differences between the two groups in CS duration used for the experiment ($F<1$), and most participants were able to complete the task with the baseline presentation time. Different fractal images were used for the main task.

Upon entering the scanner room, participants were fitted with skin conductance recording electrodes on the medial phalange of their middle and index fingers of their left hand. Electrical stimulation electrodes were then attached to both wrists. Once the participant was in position inside the scanner, a standard shock work-up procedure was conducted.

Two stimulators were used and shock amplitudes were tailored to each participant and to each wrist of every participant; set to a level that they deemed unpleasant but not painful, and of equivalent intensity for both wrists. After the shock levels were set, subjects answered the following question on a visual analogue scale (VAS): "How unpleasant do you find the shock level we have set for the experiment?" Participants could make left and right button-presses with their feet to move a cursor centered on 50 (labeled "Moderately unpleasant"), with extremes of 0 ("Not at all unpleasant") and 100 ("Extremely unpleasant"). OCD (M=60.27, SD=13.64) and Controls (M=55.46, SD=12.27) did not differ in their shock unpleasantness ratings, $F(1,68)=2.389$, $p=.127$.

Avoidance Task

One CS predicted a shock (unconditioned stimulus: US) to the participant's left wrist, while a different image predicted a shock to the participant's right wrist, in a deterministic fashion (CS+). A third CS was safe, never predicting a shock (CS-). CSs were three fractal images, whose association with left shock, right shock or safety was counterbalanced across subjects. Subjects were informed of these contingencies and experienced one trial with each CS prior to beginning the avoidance portion of the experiment. Subjects were told that they could avoid receiving shocks if they made the correct button-press response on a box placed between their feet, while a CS+ was on-screen. They were told that making a response on the right side of the box when the corresponding CS+ appeared would prevent a shock to their right wrist and similarly a left response would prevent a shock to their left wrist. These avoidance contingencies were also fully deterministic. Following these instructions, the main training stage began and consisted of 4 blocks of 30 trials (10 per CS), each lasting approximately 5 minutes. A previous study indicated that this duration was sufficient to reveal differences in habit responding between OCD patients and controls (1).

To test for habits, we used outcome devaluation technique, which relies on the fact that goal-directed behavior should be sensitive to changes in motivation, while habits are not. For all subjects, the left shock outcome was devalued by disconnecting the electrodes from participants' left wrist. The shock to the right wrist remained threating or "valued". Subjects were informed on-screen that they could no longer be shocked to the left wrist and that their only goal was to avoid receiving the remaining shock. Following this, the experimenter told participants that they would confirm the disconnection further by triggering the shocker for the left and right wrist one at a time. Subjects were asked to confirm whether or not they experienced a shock to each wrist, and all confirmed that they only received a shock to their right wrist.

Although habits and goal-directed behavior have been studied for a long time in animals, work investigating these systems in humans is recent and sparse. As humans have the capacity to verbalize, we were able to interrogate other psychological processes that might contribute to whether or not an individual displays a habit at the behavioral level. These included subjective ratings of urge to perform habits and attempts to suppress habits. We also measured subjects' expectancy of shock at various time-points, and tested their contingency knowledge after the experiment, which together controlled for the

possibility that excessive behavioral output was not a 'habit', but instead a goal-directed action driven by comprehension difficulties.

For shock expectancy ratings, the question read: "Does this stimulus currently predict a shock?". The cursor was centered on 50% ("Unsure"), where the extremes of 0 and 100% were labeled "It definitely does not" and "It definitely does", respectively. We collected these ratings six times in total, prior to each of the 4 blocks of training, immediately following the devaluation manipulation (but prior to the devaluation test- "Pre-Test") and once again, directly following the devaluation test ("Post-Test"). Pre-test ratings taken directly after the electrodes were removed served to test the efficacy of the devaluation procedure at reducing shock expectancy, and were critical in order for us to exclude the possibility that continued responding in the OCD group was not a habit, but a 'goal-directed' attempt to avoid a shock that they erroneously might have thought could still be delivered. The change from Pre to Post allowed us to test if participants' beliefs about threat were influenced by the actual performance of avoidance habits during the test stage. This is important because a previous study has shown that continued avoidance protects against the normal extinction of conditioned fear (2).

Urge to respond was quantified using a VAS rated in response to the question: "In the final section, the electrodes on the left were disconnected. Did you experience an urge to continue responding on the left side?". Subjects rated this while lying in the scanner, directly following the final expectancy rating. The cursor was centered on 50 (labeled "Moderate urge"), and the extremes were 0 ("No urge") and 100 ("Great urge"). Urge Suppression was quantified by response to a question "If you did experience an urge, did you attempt to suppress this urge?", where again the cursor was centered on 50 (labeled "Moderate effort to suppress"), with extremes of 0 ("No effort to suppress") and 100 ("Great effort to suppress"). Finally, in the contingency learning questionnaire (outside the scanner), subjects were required to identify the outcome associated with each CS (left shock, right shock, no shock) and to identify the correct avoidance response required, if any (left pedal, right pedal, none).

Image acquisition and preprocessing

fMRI data were acquired on a Siemens Magnetom Trio scanner operating 3T (Siemens Medical Solutions, Erlangen, Germany). Thirty-two interleaved transaxial sections of gradient echo, echoplanar imaging (EPI) data depicting blood oxygen level-dependent (BOLD) contrast were acquired parallel to the intercommissural line with the following parameters: repetition=2000 ms, echo time=30 ms, flip angle=78°, slice thickness=3 mm, matrix of 64x64 with FOV=192x192 mm giving 3x3 mm in-plane resolution. Prior to data analysis, the first four images were discarded for T1 equilibration. T1 structural scans were collected using magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence: 176 slices of 1 mm thickness, with TR=2300 ms, TE=2.98 ms, TI=900 ms, flip angle=9°, FOV=256x256 mm.

Data analysis

Prior to statistical analysis, imaging data were pre-processed using Statistical Parametric Mapping software (SPM8;http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). Subject data from each block were spatially realigned to the first volume in the time series of block 1. Data were then subjected to quality control tests using ArtRepair (http://cibsr.stanford.edu/tools/ArtRepair/ArtRepair.htm). Distortions arising from rapid movement (exceeding 1mm per TR) were corrected with interpolation using the average of adjacent unaffected volumes. Data were then slice-time corrected, co-registered, segmented, normalized to the Montreal Neurological Institute (MNI) template and smoothed with an 8mm Gaussian kernel. Thirty-six Controls and 44 OCD patients were initially recruited for this study, but data from 9 participants (3 controls, 6 patients) could not be included due to excessive signal dropout (n=4) or rapid motion artifacts (n=5) that could not be sufficiently corrected with interpolation. In addition, 1 patient ceased responding entirely, discovering that they were on extinction, during the habit test, and therefore their data could not be included in the analyses, leaving 33 controls and 37 OCD patients.

We conducted a *post hoc* psycho-physiological interaction (PPI) analysis analyses to test for regions showing aberrant functional connectivity with the caudate during the acquisition of avoidance. The physiological variable was the first eigenvariate of BOLD signal from our bilateral caudate ROI during early acquisition (Block 1, Warning – Safe). We used the first eigenvariate instead of the mean value across voxels, as this does not assume homogeneous responses within an area (3). The psychological variable was CS (Warning – Safe) during acquisition of avoidance, i.e. Block 1. These variables were entered as regressor in a new general linear model, in which we tested for BOLD signal that was correlated with their interaction at the first level and compared across both Study Group and Habit Group at the second level.

Skin conductance data were subjected to a high-pass filter of 0.05 hertz to remove low frequency drift, and a low-pass filter of 0.05 hertz to remove high frequency scanner noise. SCRs were defined as the baseline to peak difference within an 8 second interval following the presentation of a CS. SCR data were square-root transformed to correct for skew. As per the behavioral data, we compared SCRs to the Warning (CS+) and Safe (CS-) CSs. In the habit test, we compared SCRs to all three CSs (Devalued, Valued and Safe), to test the extent to which subjects extinguished their fear responses to the now devalued CS. SCR data from 3 subjects (2 controls and 1 patient) could not be collected due to technical difficulties. As SCR was a secondary measure, we include behavioral and brain imaging data from these subjects in all other analyses.

Behavioral Results

Habit Test

The habit data were not normally distributed, as is typical with one-shot devaluation tests (1), and therefore we conducted a Chi-square test to assess whether a greater number of subjects formed habits in the OCD group relative to controls. The number of subjects who formed habits (i.e. made any response to the Devalued CS during the habit test) was

significantly different between the Groups, χ^2=5.509, p=.019, where 41% (15/37) of OCD patients formed habits relative to 15% (5/33) of Controls. We also tested for Study Group (OCD, Control) differences in the number of false alarms to the Safe CS during the habit test; there was a non-significant trend towards a greater number of false alarms in the OCD group, F(1,69)=3.291, p=.074. There were no significant differences between Study Group (OCD, Control) or Habit Group (Habit, No Habit: as defined above) in terms of accuracy (and therefore shock experience) (Figure S1-A), false alarms (Figure S1-B) or reactions times during training. There were no differences between Habit Groups on any demographic variable (e.g. age, gender, age-of-onset) or clinical scales (e.g. anxiety).

Avoidance Training

During Training, we compared Groups (OCD, Control) on percentage of correct avoidance responses to the Warning CSs across Block (1,2,3,4). OCD patients and Controls showed equivalent performance; there was no main effect of Group, F(1,68)=1.837, p=.18, and no interaction between Group and Block, F<1 (Figure S1-A). There was a significant main effect of Block, F(3,204)=42.556, p<.001, such that accuracy improved over time. Bonferroni-corrected, pairwise comparisons revealed that accuracy was lowest in block 1 compared to each of the other blocks, all p<.001, it was also lower in block 2 compared to block 4, p<.001, and trended towards being lower in block 2 compared to 3, p=.115. There was no difference in accuracy between blocks 3 and 4, p=.221. Finally, there was no between Group (OCD, Control) difference in reactions time (RT) to the Warning CSs, no main effect of block, and no Group by Block (1,2,3,4) interaction, all F<1. This means that accuracy during training was equivalent across groups, and furthermore that OCD and control groups received the same number of shocks during training.

In addition to examining accuracy, we tested for between group differences in the percentage of commission errors made to the Safe CS across Block (1,2,3,4). Once again, there was no main effect of Group and no interaction between Group and Block, both F<1 (Figure 3B). There was a main effect of Block, F(3,204)=31.134, p<.001. As with avoidance accuracy, pairwise comparisons revealed that subjects improved their performance over time, making fewer errors of commission. There were significant differences between block 1 and blocks 2, 3 and 4, all p<.001, but no differences between the other blocks, all p>.295.

There were no differences between Habit Groups (Habit, No Habit) on avoidance accuracy or the number of false alarms, both F<1, and no interactions across Blocks, both F<1.3. There were also no differences in RT between Habit Group (Habit, No Habit), F<1, and no interaction between Habit Group and Block, F(1,35)=1.828, p=.147

Training: Skin Conductance Response (SCR)

During the training phase, we compared SCRs between Group (OCD, Control) to each CS (Warning, Safe) over Block (1,2,3,4). There was no main effect of Group and no

interaction between Group and any of the effects described below, all $F<1$ (Figure 3D). There was a significant effect of CS, such that we observed good conditioning, wherein SCRs were greater to the Warning compared to the Safe CS, $F(1,65)=235$, $p<.001$. Consistent with habituation of the SCR response over time, we found a significant main effect of Block, $F(3, 195)=16.067$, $p<.001$, such that SCRs reduced over the course of training. SCRs in block 1 were greater than in blocks 2, 3 and 4, $p<.001$. SCRs in block 2 were greater than in block 3, $p<.048$, but not block 4, $p=.191$, and blocks 3 and 4 did not differ, $p>.999$. There was a significant interaction between CS and Block, $F(3,195)=3.223$, $p=.024$, such that the difference between SCRs to stimuli was greater in Block 1 than in each of the other blocks, all $p<.05$. However, tests of simple effects confirmed that the difference between Warning and Safe CSs remained significant in each of the four blocks, all $p<.001$. We repeated this analysis, replacing Group with Habit Group (Habit, No habit), and there was no main effect of Habit Group, $F<1$, and no significant interactions between Habit Group and the within-subjects factors, all $F<1.8$.

Habit Test: Skin Conductance Response (SCR)

We compared SCRs during the habit test to the three CSs (Valued, Devalued and Safe) across Study Group (OCD, Control). There was a significant main effect of CS, $F(2,130)=16.163$, $p<.001$. Bonferroni-corrected pairwise comparisons revealed that SCRs to the Valued CS were greater than to the Devalued ($p<.002$) and Safe ($p<.001$) CSs, while the Safe and the Devalued CS did not differ significantly from one another ($p=.181$). There was no interaction between CS and Study Group, and no main effect of Study Group, both $F<1$. Together these data suggest that conditioned fear responses to the Devalued stimulus reduced in a similar manner in OCD patients and Controls (Figure 2C).

There was a significant interaction between CS and Habit Group, $F(2,68)=4.818$, $p=.011$. While the No Habit group had a significant main effect of CS, $F(2,40)=9.934$, $p=.001$, the Habit group did not, $F<1$ (Figure S1-D). Furthermore, between these groups there was no difference in SCR to the Valued CS, $F<1$, but a trend toward higher SCRs in the group of patients who formed habits to the Devalued, $F(1,34)= 3.212$, $p=.082$ and Safe CSs, $F(1,34)=2.84$, $p=.1$, relative to those who did not. There was no main effect of Habit Group $F(1,34)=1.233$, $p=.275$.

Premonitory Urge

To test if the Group difference in the urge to respond was driven by differences in habit formation, we compared OCD who did *not* form habits (n=22/37) and controls who did *not* form habits (n=27/33); the OCD subset reported a greater urge than the control subset, $U=182$, $Z=-2.586$, $p=0.01$.

The urge to respond in the OCD group was positively correlated with the OCI-r (data was missing from two patients), Spearman's $r(35)=.368$, $p=.03$, which gives greater severity values if subjects have symptoms across multiple symptom categories (e.g. washing, checking), but not with the YBOCS clinical interview, which assesses symptom severity globally independent of the spread of symptoms across categories, $r(37)=-.036$, $p=.833$,

which takes into account global severity of obsessions and compulsions, rather than the spread of severity across discrete symptom dimensions. However, the correlation with OCI-r does not survive correction for multiple comparisons and should be interpreted with caution. The urge to respond did not correlate with any of the other questionnaire data and demographic information: including state and trait anxiety, depression, verbal IQ, years in education or age of onset, all $p>.247$.

Evaluative Conditioning (Shock Expectancy)

We tested for between Study Group (OCD, Control) differences in shock expectancy to the Devalued CS. We investigated two Time Points (Pre-Test, Post-Test). Ratings at Pre-Test were made directly following the devaluation procedure, but prior to the habit test and allowed us to test the extent to which subjects understood that the devaluation procedure conferred safety. Ratings at Post-Test were collected directly following the habit test. A Mann Whitney U test did not reveal any Study Group differences in shock expectancy ratings at pre- or post-test, both $p>.35$. This indicates that OCD patients and Controls had equivalent sensitivity to the devaluation procedure in terms of reducing shock expectancy. However, we found that the Habit group had a trend towards higher expectancy ratings than the No Habit group for both Pre-Test, $U=104.5$, $Z=-2.401$, $p=.061$ and Post-Test $U=103.5$, $Z=-2.791$, $p=.056$, suggesting that there may be some association between habit formation in OCD and the extent to which the devaluation procedure was effective at reducing shock expectancy in general. This trend is not unexpected, however, as understanding of the reduction in shock likelihood is a necessary condition for showing sensitivity to devaluation. In line with this account, we repeated our main analysis of habit responding (Valued, Devalued) between groups (OCD, control) with shock expectancy to the devalued CS as a covariate, and our results were strengthened by the inclusion of this covariate. Group by CS interaction, $F(1,67)=10.163$, $p=.002$. Simple effect comparing OCD and Controls on responses to the devalued CS, $F(1,67)=13.863$, $p<.001$.

When we compared OCD patients who did *not* form habits and control subjects who did *not* form habits, there was no difference in expectancy ratings at Pre-Test, $U=269.5$, $Z=-1.078$, $p=.281$, or Post-Test, $U=289$, $Z=-.789$, $p=.43$. During training, there were trends towards greater shock expectancy to the warning CSs ($U=488$, $Z=-1.772$, $p=.076$), and lower shock expectancy to the safe CSs ($U=524$, $Z=-1.384$, $p=.166$), in the OCD group. There were no differences between the Habit and No Habit groups, both $p>.84$.

Habit Suppression

There was a trend towards a Study Group difference in the degree to which subjects reported attempting to suppress the urge to respond to the devalued CS during the habit test, $U=486$, $Z=-1.512$, $p=.078$. As above, we compared OCD and control participants who did *not* form habits, and found that these OCD patients reported greater suppression than these controls, $U(159)=-2.96$, $p=.003$. Indeed, comparing Habit and No Habit groups within the OCD cohort, the No Habit group (M=48.64, SD=30.75) reported that they attempted to suppress the urge to respond to a greater extent than the Habit group

(M=15.33, SD=28.75), *U=50, Z=-3.017*, *p*=.003, and likewise suppression was negatively correlated with the performance of habits across the entire OCD group, Spearman's *r*(37)=-.503, *p*=.002.

fMRI Results

Devalued-Safe Contrast

We caveat interpretation of the following results, but present them for the interested reader. However, in many of contrasts below (i.e. Devalued-Safe), the comparison groups differ in their behavioral output, as well as other subjectively reported experiences, such as urge to respond and attempts to actively suppress responding. Therefore there are multiple potential confounds to interpretation that must be considered. In the whole sample, comprising all OCD and Controls, there was significant activation in the right insula, left inferior parietal lobe, right supplementary motor area and right suypramarginal gyrus at the whole brain $p<.001$ uncorrected level. OCD patients showed significantly greater activation in the left middle temporal lobe compared to Controls, at $p<.001$, uncorrected. To test for activation associated with a goal-directed cessation of responding to the devalued CS, we repeated this analysis including only subjects who did not form habits (i.e., who ceased responding after devaluation), across both groups. Only activation in the left inferior parietal lobe was significant at the $p<.001$ uncorrected level across both groups, and there were no differences between OCD and Controls who did not form habits, suggesting that activation in this region is generally associated with goal-directed action control. When we compared Habit Group (Habit, No Habit) within the OCD sample on this contrast, and found that patients who formed habits had greater activation in the right supramarginal gyrus, right calcarine, left thalamus, left cerebellum and left rolandic operculum at the whole brain $p<.001$ uncorrected level. Finally, we compared Medication status (Medicated, Unmedicated) and found that unmedicated patients had greater activation compared to medicated patients in the orbital portion of the inferior frontal gyrus and the middle frontal gyrus at $p<.001$ uncorrected.

Over-Training of Avoidance in All Subjects

Across the entire sample of participants (OCD + Control), a pattern of decreasing activation across blocks was observed in the bilateral putamen, using an anatomical ROI corrected at $p<.05$ FWE. At the whole brain $p<.05$ FWE corrected level, we observed a similar decrease in activation across all subjects over time in the left supplementary motor area, $T(1,68)=5.47$ [-3,-19,52]. Regions showing this pattern at the $p<.001$, uncorrected level are presented in the supplement (Table S5). The pattern of decreasing activation over time in the putamen and SMA is consistent with studies examining automaticity and skill learning changes over time (4, 5), but not a recent study which found an increase in putamen activation over the course of over-training of an appetitive habit (6).

Supplementary Discussion

Utilization of the binary 'Habit Group' distinction in the present study was necessitated by the devaluation test. Although devaluation may be the 'gold standard' manipulation for assessing habits, future research should aim to understand the learning mechanisms that underpin this terminal behavior. A computational approach identifies contributions from 'model-based' and 'model-free' learning systems to decision-making, which putatively relate to goal-directed and habit learning, respectively (9). Using this paradigm, a recent study showed that OCD, and other disorders of compulsivity, are associated with deficits in model-based 'goal-directed' learning (10), thought to be dependent on the mOFC and caudate (10, 11). The pattern of hyper-activation we observed in this region during training in OCD patients possibly reflects abnormalities in model-based learning, to which our design was however insensitive.

Table S1. Clinical Characteristics and Demographics

	OCD	Control	df	Statistic	*p* value
N	37	33			
Gender (M:F)	18:19	19:14	1	χ^2=.558	.455
Hand (L:R)	5:32	4:29	1	χ^2=.03	.862
Smokers (Y:N)	5:32	8:24	1	χ^2=1.48	.224
Age	38.14(11.5)	37.36(12.22)	1,68	*F*=.074	.786
Education (Years)	15.54(2.55)	15.91(2.75)	1,68	*F*=.339	*.563*
Verbal IQ (NART)	111.89(7.27)	114.84(6.0)	1,61	*F*=2.928	0.092
YBOCS	21.76(6.01)	-	-	-	-
OCI-r	29.46(11.08)	5.56(4.66)	1,65	*F*=128.11	<.001
MADRS	6.92(6.68)	0.88(1.85)	1,68	*F*=25.18	<.001
YGTSS	2.43(4.78)	-	1,68	*F*=8.526	.005
STAI-State	43.89(11.54)	29.69(6.54)	1,67	*F*=37.9	<.001
STAI-Trait	55.14(13.03)	30.97(7.86)	1,67	*F*=83.58	<.001

Standard deviations are in parentheses.

NART: National Adult Reading Test; YBOCS: Yale Brown Obsessive-Compulsive Scale; OCI-r: Obsessive-Compulsive Inventory – revised; MADRS: Montgomery-Asberg Depression Rating Scale; YGTSS, Tale Global Tic Severity Scale; STAI = State Trait Anxiety Inventory.

Table S2. List of Dependent Measures, Calculation and Rationale

Measure	Calculation	Rationale
Behavior		
Avoidance Accuracy	Correct responses to warning CSs during training	Test how rapidly subjects' accuracy reached asymptote
False Alarms	Any responses to the safe CS (training and habit test)	Control for general disinhibition
Valued Stimulus	Correct responses to valued CS during habit test	Control for baseline avoidance accuracy to compare against devalued CS
Devalued Stimulus	Correct responses to devalued CS during habit test	Responding here indicates that behavior has been rendered habitual
Subjective		
Urge to Respond	VAS rating	Determine if habits, like compulsions are associated with a subjective urge to respond (premonitory urge)
Suppression	VAS rating	Determine if active behavioral suppression is a strategy employed by participants to refrain from behaving habitually
Evaluative Conditioning	Shock expectancy ratings	Control for explicit knowledge of task contingencies measured online
Contingency Test	Paper and pen test requiring participants to circle the correct response to a given fractal CS (i.e. left pedal, right pedal, none) and the outcome otherwise associated with that CS (i.e. shock, no shock)	Control for explicit knowledge of task contingencies, including action-outcome contingency, after experiment
fMRI		
Avoidance: Acquisition	Block 1 of training: Warning - Safe	Examines brain activation associated with the initial learning of avoidance
Avoidance: Over-training	Parametric modulation of activation as training progresses: (Warning-Safe)*(Blocks 1,2,3,4)	Examines brain activation as avoidance is over-trained (across blocks)
Habit Test	(OCD Habit - OCD No Habit) * (Valued-Safe)	Behavioral response to the devalued CS allow us delineate two groups, one where a habit has been formed (OCD Habit) (responses to devalued CS>0) and a group where habits have not been formed (OCD No Habit). BOLD responses to the Valued CS then reflect the online performance on habits, which has been unaffected by the devaluation manipulation (and the

	associated confounding differences in behavioral responding, urge to respond and attempts to suppress responding)

CS: conditioned stimulus; OCD: obsessive-compulsive disorder; fMRI: functional magnetic resonance imaging. BOLD: blood oxygen level dependent.

Table S3. Regions where OCD patients show hyper-activation during the acquisition of avoidance.

AAL label	Side (Left/Right)	MNI co-ordinates [x,y,z]	Dist (mm)	T	Ke
Frontal_Med_Orb	R	6,23,-11	0	4.96	128
Temporal_Mid	L	-63,-19,-8	0	4.65	75
Angular	R	42,-70,46	0	4.56	149
Temporal_Mid	L	-48,-16,-8	0	4.52	47
Temporal_Mid	R	63,-19,-17	0	4.52	27
Frontal_Inf_Oper	L	-33,17,28	1.41	4.41	24
Cingulum_Ant	L	3,44,13	0	4.38	136
Parahippocampal	L	-21,-37,-11	0	3.79	20
Precuneus	L	-3,-67,43	0	3.67	35
Cingulum_Post	R	3,-43,25	0	3.67	73
Insula	R	42,-16,4	0	3.63	12
Calcarine	R	15,-79,1	0	3.52	10
Frontal_Inf_Tri	R	45,26,-2	0	3.49	13

AAL = Automatic Anatomic Labeling. Dist = distance from AAL label. MNI = Montreal Neurological Institute. Ke = cluster size.

Study Group (OCD, Control) by CS (Warning, Safe) in Block 1 at *p*<.001 uncorrected

Table S4. Regions where OCD patients show a significant decrease in activation during over-training, but controls do not.

AAL label	Side (Left/Right)	MNI co-ordinates [x,y,z]	Dist (mm)	T	Ke
Frontal_Med_Orb	R	6,23,-11	0	4.82	76
Cingulum_Mid	L	-18,-16,46	4	4.37	26
Precuneus	R	33,-46,10	5	4.36	21
Frontal_Mid	R	33,56,19	0	4.05	16
Temporal_Mid	R	60,-46,-5	0	4.02	16
Cingulum_Mid	L	-3,-43,37	0	3.85	20
Cuneus	L	-9,-64,28	0	3.84	16
Angular	R	42,-67,34	0	3.78	27
Precuneus	R	9,-58,28	0	3.61	19

AAL = Automatic Anatomic Labeling. Dist = distance from AAL label. MNI = Montreal Neurological Institute. Ke = cluster size.

Study Group (OCD, Control) by CS (Warning, Safe) by Block (1,2,3,4) interaction at $p<.001$, uncorrected

Table S5. Regions where neural coupling between the caudate differed between Habit and No Habit groups

AAL label	Side (Left/Right)	MNI co-ordinates [x,y,z]	Dist (mm)	T	Ke
Habit > No Habit					
Olfactory (subgenual ACC: BA25)	R	[6,17,-5]	0	4.4	22
No Habit > Habit					
Inferior Frontal Gyrus	R	[48,23,-5]	0	5.11	110
Pallidum	L	[-15,2,1]	0	3.91	29

AAL = Automatic Anatomic Labeling. Dist = distance from AAL label. MNI = Montreal Neurological Institute. Ke = cluster size.

PPI analysis results at $p<.001$ uncorrected. The psychological variable was CS (Warning – Safe) in Block 1, and the physiological variable was activity in an 8mm sphere around the peak voxel in the caudate [-12,17,4] where Habit subjects showed hyperactivity relative to No Habit subjects.

Table S6. Regions that decrease in activation with over-training of instrumental avoidance in both groups

AAL label	Side (Left/Right)	MNI co-ordinates [x,y,z]	Dist (mm)	T	Ke
Supp_motor_area	L	-3,-19,52	0	5.47	602
Putamen	L	24,-1,13	0	4.82	72
Cerebellum_8	R	24,-46,-50	0	4.72	26
Occipital_mid	R	42,-73,31	0	4.41	47
Precuneus	R	30,-55,28	4.58	4.32	30
Hippocampus	L	-30,-31,-2	1	4.3	43
Putamen	R	30,2,13	0	4.29	78
Rolandic_oper	L	-39,-19,25	3.32	4.25	81
Cerebeulm 4_5	R	12,-52,-17	0	4.07	139
Occipital_mid	L	-36,-76,34	0	4.03	122
Vermis 8	-	0,-67,-32	0	4.02	40
Precentral	L	-39,-7,61	0	3.98	26
Precuneus	R	21,-64,43	1.41	3.88	55
Precuneus	L	-9,-46,7	0	3.68	21
Frontal_sup	L	-24,38,43	0	3.67	15
Cuneus	L	-18,-61,19	0	3.6	34

AAL = Automatic Anatomic Labeling. Dist = distance from AAL label. MNI = Montreal Neurological Institute. Ke = cluster size.

CS (Warning, Safe) by Block (1,2,3,4) interaction in entire sample at $p<.001$, uncorrected

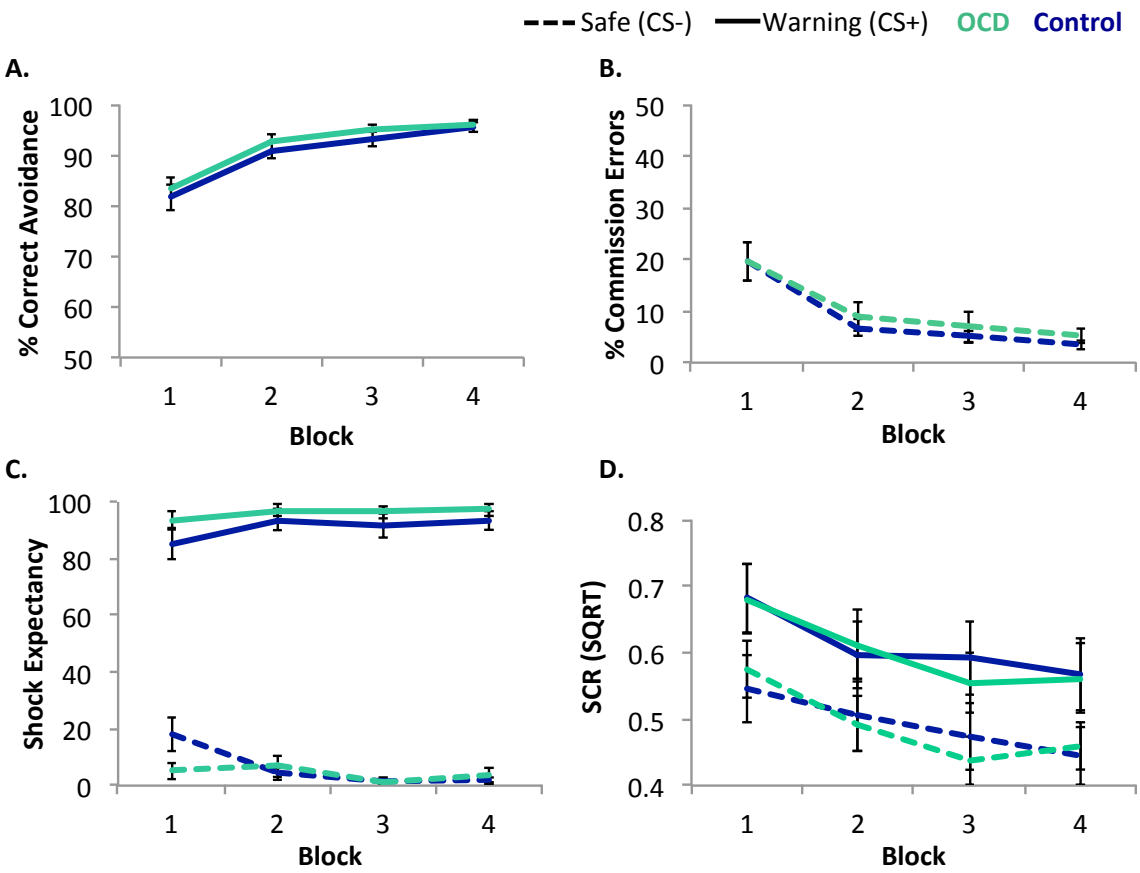

Figure S1. Training: Instrumental, Physiological and Evaluative Conditioning
Error bars denote standard error of the mean (SEM).
Panel A depicts avoidance accuracy across the 4 blocks of training to the Warning CSs. There were no differences between groups, and accuracy increased over time (see Supplement for statistic analyses).
Panel B depicts the percentage of false alarms to the Safe CS over the course of training. Like accuracy, false alarms did not differ between groups, and subjects performed better over time, i.e. made fewer false alarms (see Supplement for statistic analyses).
Panel C shows the evaluative conditioning performance of subjects, i.e. shock expectancy ratings over the course of training. There were trends towards greater shock expectancy to the warning CSs (U=488, p=.076), and lower shock expectancy to the safe CSs (U=524, p=.166), in the OCD group (n=37) compared to Controls (n=33).
Panel D illustrates physiological conditioning measured using skin conductance responses (SCRs); OCD patients (n=36) and Controls (n=31) showed equally good differentiation between the CS+ and CS-, and there was a general habituation effect over time in both groups (note 2 controls and 1 patient were excluded from this analysis due to insufficient SCR data).

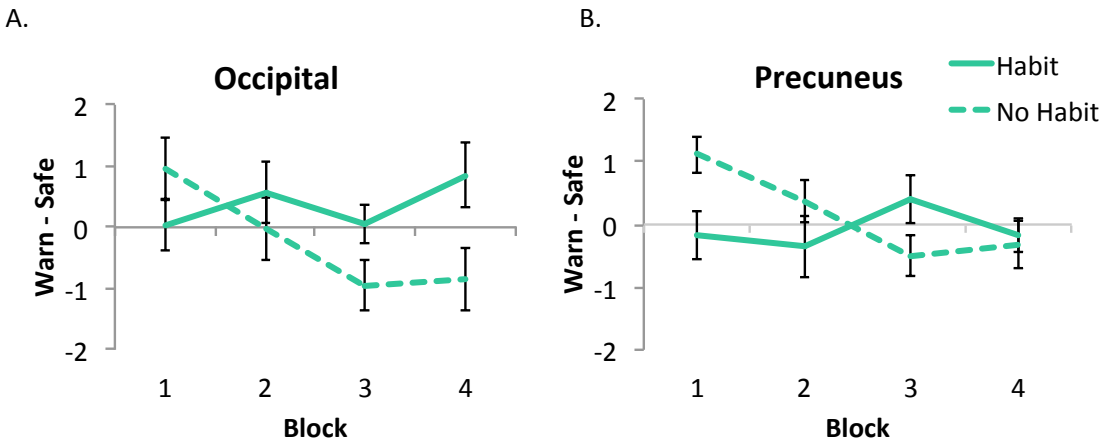

Figure S2. Differences between Habit and No Habit groups during over-training at *p*<.001 uncorrected.
Panel A depicts a significant interaction between Habit Group in the right occipital gyrus *T*(1,35)=4.42, Z=3.91 ([27,-94,16], Ke=22).
Panel B depicts a significant interaction between Habit Group in the right precuneus *T*(1,35)=4.11, Z=3.68 ([18,-49,34], Ke=14)

Supplementary References

1. Gillan CM*, et al.* (2014) Enhanced avoidance habits in obsessive-compulsive disorder. *Biol Psychiatry* 75(8):631-638.
2. Lovibond PF, Mitchell CJ, Minard E, Brady A, & Menzies RG (2009) Safety behaviours preserve threat beliefs: Protection from extinction of human fear conditioning by an avoidance response. *Behav Res Ther* 47(8):716-720.
3. Friston KJ, Rotshtein P, Geng JJ, Sterzer P, & Henson RN (2006) A critique of functional localisers. *Neuroimage* 30(4):1077-1087.
4. Poldrack RA*, et al.* (2005) The neural correlates of motor skill automaticity. *J Neurosci* 25(22):5356-5364.
5. Ashby FG, Turner BO, & Horvitz JC (2010) Cortical and basal ganglia contributions to habit learning and automaticity. *Trends Cogn Sci* 14(5):208-215.
6. Tricomi E, Balleine BW, & O'Doherty JP (2009) A specific role for posterior dorsolateral striatum in human habit learning. *Eur J Neurosci* 29(11):2225-2232.
7. Aron AR, Robbins TW, & Poldrack RA (2004) Inhibition and the right inferior frontal cortex. *Trends Cogn Sci* 8(4):170-177.
8. Liljeholm M, Tricomi E, O'Doherty JP, & Balleine BW (2011) Neural Correlates of Instrumental Contingency Learning: Differential Effects of Action-Reward Conjunction and Disjunction. *Journal of Neuroscience* 31(7):2474-2480.
9. Daw ND, Niv Y, & Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* 8(12).
10. Voon V*, et al.* (in press) Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry*.
11. Otto A, Gershman S, Markman A, & Daw N (2013) The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychological Science* 24(5):751-761.