Aversive Pavlovian control of instrumental behavior in humans

Geurts, Dirk E M <javascript:contributorCitation('Geurts, Dirk E M');>; Huys, Quentin J M <javascript:contributorCitation('Huys, Quentin J M');>; den Ouden, Hanneke E M <javascript:contributorCitation('den Ouden, Hanneke E M');>; Cools, Roshan <javascript:contributorCitation('Cools, Roshan');>

Abstract: Adaptive behavior involves interactions between systems regulating Pavlovian and instrumental control of actions. Here, we present the first investigation of the neural mechanisms underlying aversive Pavlovian-instrumental transfer using fMRI in humans. Recent evidence indicates that these Pavlovian influences on instrumental actions are action-specific: Instrumental approach is invigorated by appetitive Pavlovian cues but inhibited by aversive Pavlovian cues. Conversely, instrumental withdrawal is inhibited by appetitive Pavlovian cues but invigorated by aversive Pavlovian cues. We show that BOLD responses in the amygdala and the nucleus accumbens were associated with behavioral inhibition by aversive Pavlovian cues, irrespective of action context. Furthermore, BOLD responses in the ventromedial prefrontal cortex differed between approach and withdrawal actions. Aversive Pavlovian conditioned stimuli modulated connectivity between the ventromedial prefrontal cortex and the caudate nucleus. These results show that action-specific aversive control of instrumental behavior involves the modulation of fronto-striatal interactions by Pavlovian conditioned stimuli.

DOI: https://doi.org/10.1162/jocn_a_00425

Aversive Pavlovian Control of Instrumental Behavior in Humans

Dirk E. M. Geurts[1], Quentin J. M. Huys[2,3], Hanneke E. M. den Ouden[1], and Roshan Cools[1]

Abstract

■ Adaptive behavior involves interactions between systems regulating Pavlovian and instrumental control of actions. Here, we present the first investigation of the neural mechanisms underlying aversive Pavlovian–instrumental transfer using fMRI in humans. Recent evidence indicates that these Pavlovian influences on instrumental actions are action-specific: Instrumental approach is invigorated by appetitive Pavlovian cues but inhibited by aversive Pavlovian cues. Conversely, instrumental withdrawal is inhibited by appetitive Pavlovian cues but invigorated by aversive Pavlovian cues. We show that BOLD responses in the amygdala and the nucleus accumbens were associated with behavioral inhibition by aversive Pavlovian cues, irrespective of action context. Furthermore, BOLD responses in the ventromedial prefrontal cortex differed between approach and withdrawal actions. Aversive Pavlovian conditioned stimuli modulated connectivity between the ventromedial prefrontal cortex and the caudate nucleus. These results show that action-specific aversive control of instrumental behavior involves the modulation of fronto-striatal interactions by Pavlovian conditioned stimuli. ■

INTRODUCTION

Adaptive behavior depends on interactions between systems regulating affective versus rational, instrumental control (Huys et al., 2011; Evans, 2008; Daw, Niv, & Dayan, 2005). Many decision-making phenomena that appear irrational, such as the framing effect (Tversky & Kahneman, 1981) and the optimism bias (Sharot, Riccardi, Raio, & Phelps, 2007; Weinstein, 1980), may reflect Pavlovian impact of affective cues on instrumental behavior (Dayan & Huys, 2008; Dayan, Niv, Seymour, & Daw, 2006). Elucidating the neural mechanisms underlying Pavlovian effects on instrumental actions is crucial, not just for understanding normal behavior but also because Pavlovian effects are implicated in neuropsychiatric disorders (e.g., addiction and depression; Flagel et al., 2011; Dayan & Huys, 2008). Here we investigate these mechanisms by using fMRI and a well-established paradigm for assessing Pavlovian influences on instrumental responding: Pavlovian–instrumental transfer (PIT).

Existing neuroimaging work on PIT has focused on the potentiation of appetitive instrumental responding by appetitive cues (Bray, Rangel, Shimojo, Balleine, & O'Doherty, 2008; Talmi, Seymour, Dayan, & Dolan, 2008). For example, Talmi et al. (2008) have revealed BOLD responses in the nucleus accumbens and the amygdala during appetitive PIT. However, no imaging study and only a few behavioral studies have addressed the effects of aversive cues on human behavior (Huys et al., 2011; Di Giusto, Di Giusto, & King, 1974). This is pertinent, because the influence of aversive expectations on behavior likely plays an important role in several psychiatric conditions (Bijttebier, Beck, Claes, & Vandereycken, 2009).

We adapted a paradigm that previously showed significant behavioral PIT of both appetitive and aversive cues (Huys et al., 2011). Our first question was whether structures identified as contributing to appetitive PIT—amygdala and nucleus accumbens—are also involved in aversive PIT. The second question concerned action specificity, an aspect of PIT that so far has received little attention. We have recently discovered that the effect of Pavlovian cues depended on the valence of instrumental behaviors: Whereas appetitive Pavlovian conditioned stimuli (CSs) potentiated approach and inhibited withdrawal, aversive CSs suppressed approach (as in conditioned suppression) but potentiated withdrawal (Huys et al., 2011). This finding resonates with the fact that many neuropsychiatric disorders prominently involve abnormal control not only of appetitive behaviors (e.g., approach) but also of aversive behaviors (e.g., withdrawal; Trew, 2011). If Pavlovian cues have opposite effects on these different actions, then a better understanding of the mechanisms underlying this action specificity should help resolve how instrumental behavior is controlled by Pavlovian cues.

Action specificity suggests that affective cues might interact differently with systems that code for approach or

[1]Radboud University Nijmegen Medical Centre, The Netherlands, [2]University College London, [3]Guy's and St. Thomas' National Health Service Foundation Trust, London, UK

Journal of Cognitive Neuroscience 25:9, pp. 1428–1441
doi:10.1162/jocn_a_00425

withdrawal. We asked whether action specificity in Pavlovian control involves differential influences on neural regions that encode action specificity. One possibility is that it involves direct Pavlovian modulation of regions that encode action specificity. Another possibility is that Pavlovian cues modulate the influence of regions that are action specific on regions that implement instrumental behavior. One region prominently associated with instrumental behavior is the striatum (the caudate nucleus and putamen; Balleine & O'Doherty, 2010). We tested these hypotheses by conducting univariate analyses of action-specific PIT effects as well as functional connectivity analyses of action-specific influences on the striatum during PIT.

METHODS

Participants

Fifteen right-handed volunteers participated in a behavioral experiment conducted in a dummy scanner environment before the fMRI experiment ("behavioral group"). Subsequently, 20 right-handed volunteers participated in the fMRI experiment ("fMRI group"). The experiment was approved by the local ethics committee. Exclusion criteria were claustrophobia, neurological or cardiovascular diseases, psychiatric disorders, regular use of medication, use of psychotropic drugs, smoking, or metal parts in the body. Written informed consent was obtained before study procedures. Two fMRI participants were removed from analyses because of below-chance performance in the final stage of the instrumental learning phase and/or during the Pavlovian query trials. For two other fMRI participants, one of the two sessions was excluded: one participant did not complete the first session because of discomfort in the scanner, and the juice delivery setup failed for another participant's first session. Accordingly, data are reported from 15 participants (six women; mean age = 25.7 years, SD = 3.4 years) in the behavioral group and 18 participants (11 women; mean age = 23.8 years, SD = 3.5 years) in the fMRI group.

Pavlovian–Instrumental Transfer Paradigm

Participants performed the experimental task adapted from Huys et al. (2011). The paradigm was programmed using Matlab (2009b, TheMathWorks, Natick, MA) with the Psychophysical Toolbox extension (Brainard, 1997). The experiment consisted of two sessions, each with three stages: (i) instrumental training, (ii) Pavlovian conditioning, and (iii) PIT. The setup of the experiment was the same for the two sessions, but different instrumental and Pavlovian stimuli were used in each session.

Two major adaptations were made to the version used by Huys et al. (2011): First, unlike Huys et al., primary outcomes (juices) were used for Pavlovian conditioning, whereas secondary outcomes (monetary) were used for instrumental training. This was done to make sure that the (de)motivating effects of the Pavlovian CSs were not because of similarity in outcome with the instrumental action. This made our paradigm sensitive to general as opposed to outcome-specific motivating effects of Pavlovian CSs. Second, participants had to press a button multiple times rather than just once. This generated an additional dependent variable (the number of button presses), which we anticipated to be sensitive to PIT (cf. Talmi et al., 2008) and allowed us to look at parametric PIT effects in our fMRI analysis.

Instrumental Training

The instrumental task (Figure 1A) was framed in terms of an approach/withdrawal go/no-go task. On each trial, an instrumental stimulus (mushroom or shell) was presented centrally at the top of the screen. A dot appeared at the bottom of the screen and moved upward at a constant speed (reaching the top in 2.5 sec). Participants had to choose whether to collect the instrumental stimulus by steering the dot through it or whether not to collect it by steering it past the stimulus. Each choice resulted in monetary wins or losses (±5 cents). Participants influenced the trajectory of the dot by pressing one button repeatedly. Every button press added a fixed sideways displacement to the dot trajectory. This displacement decayed back to zero over time at a speed that was calibrated before the experimental session to the maximum frequency at which participants were able to press the button (mean maximum frequency was 4.9 Hz, SD = 2.0). There were two action contexts: In the approach context, the dot appeared in one of the bottom corners and, in the absence of button presses, moved past the instrumental stimulus. Thus, participants had to actively press the button repeatedly to move the dot centrally toward the instrumental stimulus and collect it. In the withdrawal context, the dot appeared in the middle of the screen and by default moved upward through the instrumental stimulus. In this case, button presses were required to move the dot away from the instrumental stimulus to avoid collecting it. Thus, there were four trial types: approach-go, approach-no-go, withdrawal-go, and withdrawal-no-go (Figure 1A). Thus the Action Context determined whether the active response was an approach or a withdrawal response. Whether an instrumental stimulus was collected was determined based on whether the dot entered a goal region (invisible to the participant; Figure 1A) around the instrumental stimulus. If the dot entered this region (after go-approach or no-go-withdrawal), then the stimulus was collected. If not (after no-go-approach or go-withdrawal), then the stimulus was not collected. At times, the dot could touch the target area on the side, only entering it partially. In this case, feedback consisted of the words: "pressed, but incomplete action" and no money was won or lost. At the end of each full action, monetary feedback ("+5 cents" or "−5 cents") was displayed.

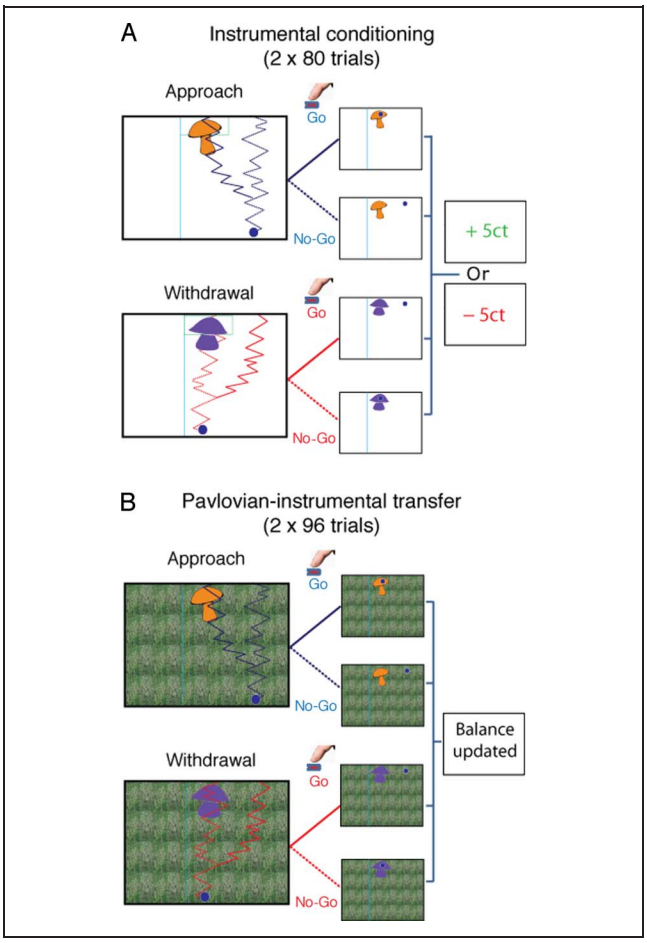

Figure 1. Task description. (A) Instrumental training. Trials started with the appearance of the instrumental stimulus at the top center of the screen and of a dot at the bottom of the screen. In approach trials, the dot started either on the left or on the right bottom side of the screen. Participants could choose to do nothing (approach-no-go), in which case the dot would wiggle past the instrumental stimulus. Alternatively, they could push the button repeatedly to steer the dot through the instrumental stimulus (approach-go). In withdrawal trials, the dot started centrally at the bottom beneath the instrumental stimulus. Participants could choose to push the button repeatedly to avoid moving through instrumental stimulus (withdrawal-go) or to do nothing (withdrawal-no-go). The four possible trajectories are drawn in the figure (red and blue lines). The green square around the stimulus (invisible to the participant) was the goal region. If the dot entered the goal region, then the instrumental stimulus was collected. The straight line just to one side of the instrumental stimulus was a reflecting boundary that the dot could not cross. Timings were as follows: Instrumental stimuli were presented for 2.5 sec, during which responses were collected. After 2.5 sec, feedback was presented for 1 sec. The ITI was 1 sec (blank screen). (B) PIT. This paralleled the instrumental training, except that Pavlovian stimuli tiled the background. No outcomes were presented, but participants were instructed that their choices counted toward the final total. Participants were explicitly instructed that the juices were collected outside the scanner, and they agreed before the start of the experiment to drink them afterward. Timing of one trial was as follows: 250 msec after the onset of the Pavlovian stimulus, the instrumental stimulus (and dot) was overlaid on top of this Pavlovian stimulus. Duration of the instrumental stimulus was 2.5 sec; duration of the Pavlovian stimulus was 2.75 sec. Upon offset of both stimuli, feedback was presented, which consisted only of the words "Balance is updated" (duration = 1 sec, ITI = 1 sec).

To orthogonalize the approach-withdrawal and appetitive-aversive axes, the learned instrumental values in approach and withdrawal blocks needed to be matched. To achieve this, both go and no-go responses were, if correct, rewarded to the same extent. Additionally, to avoid a confound of behavioral activation, in each condition (i.e., in both approach and withdrawal conditions) the go action was designated as the correct response for half of the instrumental stimuli, and the no-go action for the other half. Incorrect responses had opposite outcome contingencies to correct responses, yielding more punishments than rewards. This ensured that go, no-go, approach, and withdrawal overall had the same learned association with rewards and punishments. In both the approach and withdrawal context, there were two go stimuli, which yielded reward more often after active responses (and punishment after not responding), and two no-go stimuli, which yielded reward more often after not responding (and punishment after go responding). Reinforcement was probabilistic with probabilities ranging from 0.6 to 1 (on average, the ratio reward/punishment following a correct action was 0.85:0.15 for go stimuli and 0.8:0.2 for no-go stimuli; the difference arose from a technical error). Trials were labeled as correct if participants chose the usually rewarded response.

Average reinforcement was matched between approach and withdrawal contexts (behavioral group: mean proportion of positively reinforced trials for approach = 0.58; for withdrawal = 0.61, paired sample t test: $t(14) = -0.8, p = .4$; fMRI group: mean proportion of positively reinforced trials for approach = 0.63; for withdrawal = 0.64; paired sample t test: $t(17) = -0.14, p = .9$). Accordingly, the difference between approach and withdrawal actions cannot be driven by Pavlovian responses to the instrumental stimuli. Thus, rather than representing effects of competing Pavlovian responses, the effects we report represent PIT effects.

Every session consisted of 80 instrumental training trials alternating between blocks of eight approach and eight withdrawal trials. Initial stimuli and action context were randomized across participants.

Pavlovian Conditioning

Each Pavlovian conditioning trial started with the presentation of one of three audiovisual stimuli consisting of a pure tone and a fractal. The appetitive and aversive Pavlovian CSs were followed, respectively, by 2 ml of appetitive or aversive juice (i.e., the unconditioned stimuli [US]) on 50% of trials. The neutral CS was followed by no (juice) outcome. Before the fMRI experiment, participants indicated their preference for apple juice, orange juice, or strawberry lemonade. The aversive juice was a bitter solution of magnesium sulphate (0.3 M). Each Pavlovian CS was presented 20 times, and for each session there was a separate set of three stimuli. Stimulus presentation order was fully randomized across participants. Stimulus duration was 4.5 sec, and juice delivery occurred between 0.5 and

1.5 sec after stimulus onset. The intertrial interval (ITI) was 1 sec.

To test and stimulate task involvement during conditioning, query trials were presented after every 10 Pavlovian trials. On these trials, participants chose one of the two presented Pavlovian stimuli (presented for 2 sec; ITI 0.5 sec) in extinction, that is, there were no outcomes in these trials. The outcomes were only recorded for the last session (because of technical error). In the fMRI group we further assessed conditioning by asking participants to indicate the degree to which they liked each of the juices and the Pavlovian CSs by means of visual analogue scales (VAS), before and after the experiment.

Pavlovian–Instrumental Transfer

Stimulus presentation was the same as in the instrumental training stage, except that (i) Pavlovian stimuli tiled the background from 250 msec before and during the instrumental trial and (ii) no monetary feedback and no juice outcomes were presented (Figure 1B). However, participants were instructed that their choices counted toward the final monetary total and that the juices associated with the Pavlovian stimuli were collected outside the scanner for them to drink afterwards, that is, PIT was conducted in nominal extinction.

Participants performed 96 PIT trials per session, alternating between miniblocks of eight approach and eight withdrawal trials. Initial instrumental stimulus, CS, and action context were randomized. The numbers of go and no-go stimuli were matched between conditions (i.e., Action Context × CS Valence). After every trial, feedback consisted of a screen displaying "Balance is being updated."

Image Acquisition

Whole-brain imaging was performed on a 3-Tesla MR scanner (Magnetrom Tim Trio, Siemens Medical Systems, Erlangen, Germany). Functional data were obtained using a multiecho gradient T2*-weighted EPI (ME-EPI) scanning sequence (Poser, Versluis, Hoogduin, & Norris, 2006) with BOLD contrast (38 axial-oblique slices; repetition time = 2.32 sec; echo times = 9.0, 19.3, 30, and 40 msec; in plane resolution = 3.3 × 3.3 mm; slice thickness = 2.5 mm; distance factor = 0.17; flip angle = 90°). Visual stimuli were projected on a screen and were viewed through a mirror attached to the head coil. In addition, a high-resolution T1-weighted magnetization-prepared rapid-acquisition gradient-echo anatomical scan was obtained from each participant (192 sagittal slices; repetition time = 2.3 sec; echo time = 3.03 msec; voxel size = 1.0 × 1.0 × 1.0 mm; field of view = 256 mm).

Behavioral Data Analysis

The behavioral data were analyzed using the statistic software SPSS 16.0, and the modeling was performed in Matlab (2009b).

Instrumental Training

First, we assessed change in performance over time during instrumental training. The proportion of correct responses was calculated for the first eight and last eight trials separately for each of the four trial types. To assess whether participants learned to make the correct choice during instrumental training, data were averaged across sessions and submitted to a repeated-measures ANOVA with Time Bin (two levels: beginning/end of instrumental training), Action Context (two levels: approach/withdrawal), and Response Type (two levels: go/no-go) as within-subject factors and Group (two levels: behavioral/fMRI) as between-subject factor. Second, we assessed whether the learned behavior generalized to and over the PIT stage. This was done with the same ANOVA with the difference that the factor Time Bin was changed to include three levels: the end of the instrumental training and the beginning and the end of the PIT stage.

Pavlovian Conditioning

To assess Pavlovian conditioning, we investigated whether the proportion of correct choices on query trials differed from chance. In addition, liking ratings of the CSs before and after conditioning were analyzed using an ANOVA with Time of Rating (two levels: before/after conditioning) and Valence (three levels: appetitive/neutral/aversive) as within-subject factors.

Pavlovian–Instrumental Transfer

There were two dependent measures: choice (go/no-go) and the number of button presses on go trials. Go trials were defined as those PIT trials on which one or more than one button press was made. All behavioral outcome measures were averaged across sessions and submitted to ANOVAs with Action Context (two levels: approach/withdrawal), and Pavlovian CS Valence (three levels: appetitive/neutral/aversive) as within-subject factors and Group (two levels: behavioral/fMRI) as between-subject factor. Planned contrasts were targeted at effects of aversive PIT, that is, the primary focus of this study. For these follow-up analyses, the three-level factor Pavlovian CS Valence in the omnibus ANOVA was replaced by a Pavlovian CS Valence factor with two levels: aversive and neutral.

Model-based Analyses

We anticipated that the expectation associated with each instrumental stimulus would contribute to the BOLD response. Therefore, we computed these expectations (so-called instrumental Q values) using a reinforcement learning model and included them in the fMRI analysis. The reinforcement learning model and the fitting procedures are described in detail in Huys et al. (2011). After

fitting the parameters, the action values $Q_{t1}(a_t; s_t)$ determining choice probabilities on trial t were extracted and used in the fMRI analysis.

fMRI Analysis

fMRI data analysis was performed with SPM5 software (Statistical Parametric Mapping; Wellcome Trust Centre for Cognitive Neuroimaging, London, UK). The first five volumes of each participant's data set were discarded to allow T1 equilibrium.

First, realignment parameters were estimated for the images acquired at the first echo time and consequently applied to images resulting from the three other echoes. The echo images were combined by applying a PAID-weight algorithm assessing the signal-to-noise ratio as described by Poser et al. (2006). Thirty volumes, acquired before each instrumental training session, were used as input for this algorithm. Thereafter, the following preprocessing steps were applied: slice-time correction, coregistration, and a segmentation procedure using the tissue probability maps provided by SPM5 for gray matter, white matter, and CSF centered in Montreal Neurological Institute (MNI) space to estimate normalization parameters based on the structural image. Structural as well as functional images were then normalized by applying these estimations. All normalized images were smoothed with an isotropic 8 mm FWHM Gaussian kernel (Worsley & Friston, 1995).

A random effects, event-related, statistical analysis was performed with SPM5. This analysis was restricted to the PIT stage. First, we specified a separate general linear model (GLM) for each participant (Figure 2). For each session, six main regressors represented the six PIT trials: (1) approach appetitive, (2) approach neutral, (3) approach aversive, (4) withdrawal appetitive, (5) withdrawal neutral, and (6) withdrawal aversive. For each main regressor, two additional parametric regressors were added (Büchel, Wise, Mummery, Poline, & Friston, 1996): (i) One regressor represented the tonic parametric modulation of BOLD responses during each trial by the number of button presses per trial: the PIT regressor (cf. Talmi et al., 2008). (ii) Another regressor represented the parametric modulation of BOLD responses by the Q value per trial as estimated from the model-based analysis. These parametric modulators were serially orthogonalized. Additionally, a regressor of no interest modeled phasic button presses as single events (cf. Talmi et al., 2008). All paradigm-related regressors were modeled as delta functions at the onset of the instrumental stimulus presentation per trial and were convolved with a canonical hemodynamic response function. Realignment parameters (the three rigid-body translations and three rotations) were added to capture residual movement-related artifacts. High-pass filtering (128 sec) was applied to the time series of the functional images to remove low-frequency drifts. Parameter estimates for all regressors were obtained by maximum-likelihood estimation, modeling temporal autocorrelation (AR1). The parameter estimates, derived from this fit of the model to

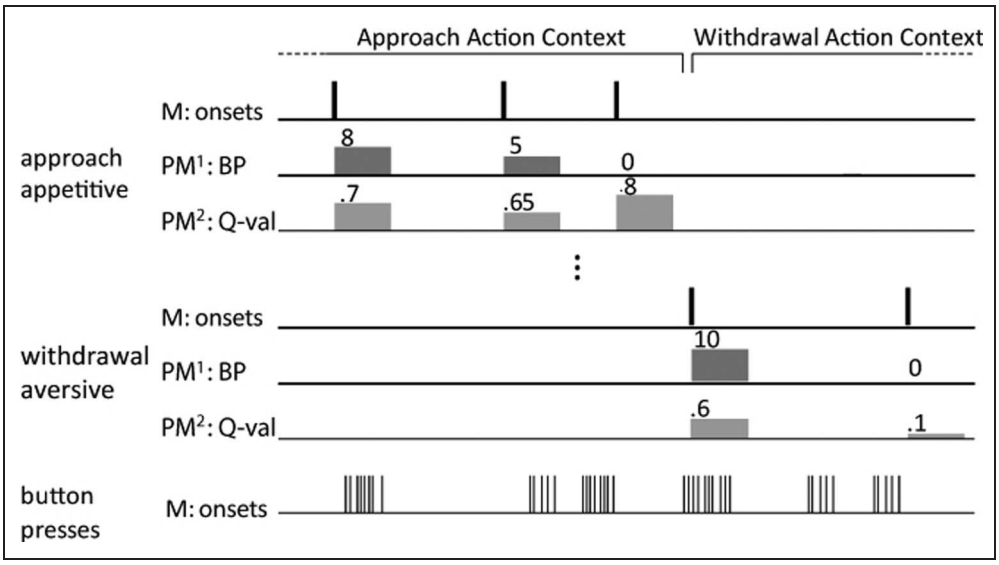

Figure 2. Schematic depiction of the GLM to analyze the PIT data (figure after Talmi et al., 2008). The main regressors (M) model the onset of a trial as a delta function. There is a main regressor for each of the six trial types. For all six main regressors, there are two parametric modulators (PM). The first parametric modulater (PM^1), the PIT regressor, consists of the number of button presses made per trial (0 for no-go). The second parametric modulator (PM^2) represents the Q value for each chosen action dependent on the instrumental stimulus shown in the trial at hand. In the seventh main regressor (of no interest), every single button press is modeled by a delta function. For reasons of clarity, two of the six trial types (approach appetitive and withdrawal aversive) are depicted only for one session and no movement nuisance regressors are shown.

the data, reflect the strength of covariance between the data and the canonical response function for each of the regressors.

Parameter estimates for the six parametric PIT regressors were estimated at the subject-level and then used in a 2 × 3 ANOVA (full factorial design) at the group level with factors Action Context (two levels: approach/withdrawal) and CS Valence (three levels: appetitive/neutral/aversive) as within-subject factors. Restricted maximum likelihood estimates of variance components were used to allow for unequal variance between subjects and possible deviations from sphericity introduced by dependencies between levels in the repeated-measures design. The main effects and interactions were then calculated. We assessed the following three planned contrasts to test our hypotheses, which focused on aversive PIT:

1. Main effect of CS Valence, contrasting aversive and neutral CSs ([approach neutral + withdrawal neutral] − [approach aversive + withdrawal aversive]). This contrast identified CS-dependent coupling of the BOLD response with the number of button presses, that is, aversive PIT-related BOLD responses independent of Action Context.
2. Interaction between Action Context and CS Valence ([approach neutral − approach aversive] − [withdrawal neutral − withdrawal aversive]). This contrast identified BOLD responses associated with action-specific aversive PIT.
3. Main effect of Action Context, contrasting approach and withdrawal trials ([approach appetitive + approach neutral + approach aversive] − [withdrawal appetitive + withdrawal neutral + withdrawal aversive]). This contrast identified regions where BOLD responses are action specific, that is, differ between approach and withdrawal.

To investigate the valence specificity of the effects, supplementary analyses were conducted to assess the same three contrasts, with the CS Valence factor contrasting appetitive with aversive CSs and appetitive with neutral CSs.

It is important to note that the parametric nature of the PIT regressor ensures that the contrasts of interest represent BOLD response involved in PIT and do not reflect differences in motor activity or Pavlovian CS per se (cf. Talmi et al., 2008). However, this analysis explicitly discounts signals that are constant, that is, do not vary as a function the number of button presses during the presentation of each CS. Therefore, following Talmi et al. (2008), to take such signals into account we also contrasted the main regressors (instead of the parametric PIT regressors) at the subject level to calculate both a main effect of CS Valence ([approach neutral + withdrawal neutral] − [approach aversive + withdrawal aversive]) and an interaction between CS Valence and Action Context ([approach neutral − approach aversive] − [withdrawal neutral − withdrawal aversive]). The resulting statistical parametric maps for each contrast were then used to conduct a *t* test at the group level with behavioral aversive PIT effects as a covariate. The behavioral aversive PIT effect for each subject was computed in terms of the average number of button presses, irrespective of Action Context ([approach neutral + withdrawal neutral] − [approach aversive + withdrawal aversive]), and as a function of Action Context ([approach neutral − approach aversive] − [withdrawal neutral − withdrawal aversive]). These analyses revealed regions in which CS-dependent BOLD responses were associated with individual behavioral PIT effects.

Functional Connectivity Analyses

Next we assessed whether action specificity of behavioral aversive PIT was accompanied by action-specific PIT-related functional connectivity. Specifically, we conducted psychophysiological interaction (PPI) analysis to assess whether action-specific PIT was associated with PIT-related modulation of functional connectivity with seed regions exhibiting a main effect of Action Context. First, for each individual, the (first principal component of the) BOLD time series was extracted from an 8-mm sphere surrounding the BOLD response peak revealed by the main Action Context contrast (the seed; the ventromedial prefrontal cortex [vmPFC]; Figure 6). The time series was then deconvolved based on the canonical hemodynamic response model to construct a time series of neural BOLD responses following the procedures outlined by Gitelman, Penny, Ashburner, and Friston (2003). Second, for every participant, two GLMs were estimated, one for each Action Context, which included the following three regressors (as well as the six motion parameters): (1) the seed BOLD response time series; (2) a parametric task contrast regressor representing aversive PIT (neutral minus aversive); and (3) the PPI regressor, that is, the interaction between (1) and (2), computed by multiplication of the deconvoluted regressor (1) and regressor (2). The PPI regressor was then convolved with the hemodynamic response function. Parameter estimates for the PPI regressor were estimated by maximum-likelihood estimation, modeling temporal autocorrelation (AR1) at the subject level, and were then used in a *t* test at the group level. The parameter estimates, derived from this fit of the model to the data, reflect the strength of PIT-related connectivity with the action-specific seed region (the vmPFC). To assess the relationship between individual behavioral PIT effects and functional PIT-related connectivity, covariates representing behavioral PIT effects (average number of button presses during neutral minus aversive trials) were included in the second level group analysis.

Statistical Thresholding and Volumes of Interest

We report only those effects that survive family-wise error (FWE) correction for multiple comparisons at the whole brain ($p_{\text{FWE WB}} < .05$, voxel level) or within volumes of interest ($p_{\text{FWE SV}} < .05$, voxel level). On the basis of

existing literature (Corbit & Balleine, 2005, 2011; Talmi et al., 2008), we expected PIT effects in the amygdala and the nucleus accumbens. Therefore, these regions were defined as volumes of interest, using anatomical criteria. The bilateral amygdala was defined using the automated anatomical labeling atlas (Tzourio-Mazoyer et al., 2002). The bilateral nucleus accumbens was segmented for each participant using the FSL FIRST segmentation tool (Patenaude, Smith, Kennedy, & Jenkinson, 2011). These individual segments were then overlaid onto each other, generating one nucleus accumbens for the group. The amygdala and accumbens volumes were combined, so that voxel level correction for multiple comparisons was conducted for all voxels within these two volumes. Furthermore, we had a specific hypothesis regarding the action specificity of the PIT effects. In particular, we reasoned that action specificity of PIT might arise from Pavlovian effects on neural regions known to implement instrumental action. One of the most prominent regions implicated in instrumental action control is the striatum (Balleine & O'Doherty, 2010). Therefore, we conducted additional (univariate and connectivity) analyses of action-specific effects in the bilateral striatum, defined as the caudate nucleus and putamen based on the automated anatomical labeling atlas (Tzourio-Mazoyer et al., 2002).

RESULTS

Behavioral Data

Behavioral data are reported across the behavioral (n = 15) and fMRI group (n = 18). To facilitate interpretation of the fMRI results, we additionally present the data for the fMRI group separately. However, no significant differences between the groups were found.

Instrumental Conditioning

Analysis of the first stage of the experiment indicates robust instrumental learning (Figure 3A). Participants learned to make correct choices during the instrumental learning stage indicated by an increasing number of correct responses over time, $F(1, 31) = 97.9$, $p < .001$. Furthermore, these differences were affected by the action context (approach or withdrawal) and changed over time: There was a significant three-way interaction between Time Bin, Action Context, and Response Type, $F(1, 31) = 34.0$, $p < .001$. This was because of participants initially preferring to approach the instrumental stimulus, that is, to go during approach (approach-go vs. no-go at the beginning of instrumental training: $t(32) = 8.9$, $p < .001$) but to no-go during withdrawal (withdrawal-go vs. no-go: $t(32) = -4.6$, $p < .001$; simple interaction effect between Action Context and Response Type at the end of instrumental training: $F(1, 31) = 87.6$, $p < .001$). The bias toward withdrawal-no-go disappeared and the bias toward approach-go became less strong but remained significant during learning (withdrawal-go vs. no-go at the end of instrumental training: $t(32) = -1.4$, $p > .1$; appoach-go vs. no-go at the end of instrumental training: $t(32) = 2.4$, $p < .05$; simple interaction effect between Action Context × Response Type at the end of instrumental training: $F(1, 31) = 7.1$, $p < .05$).

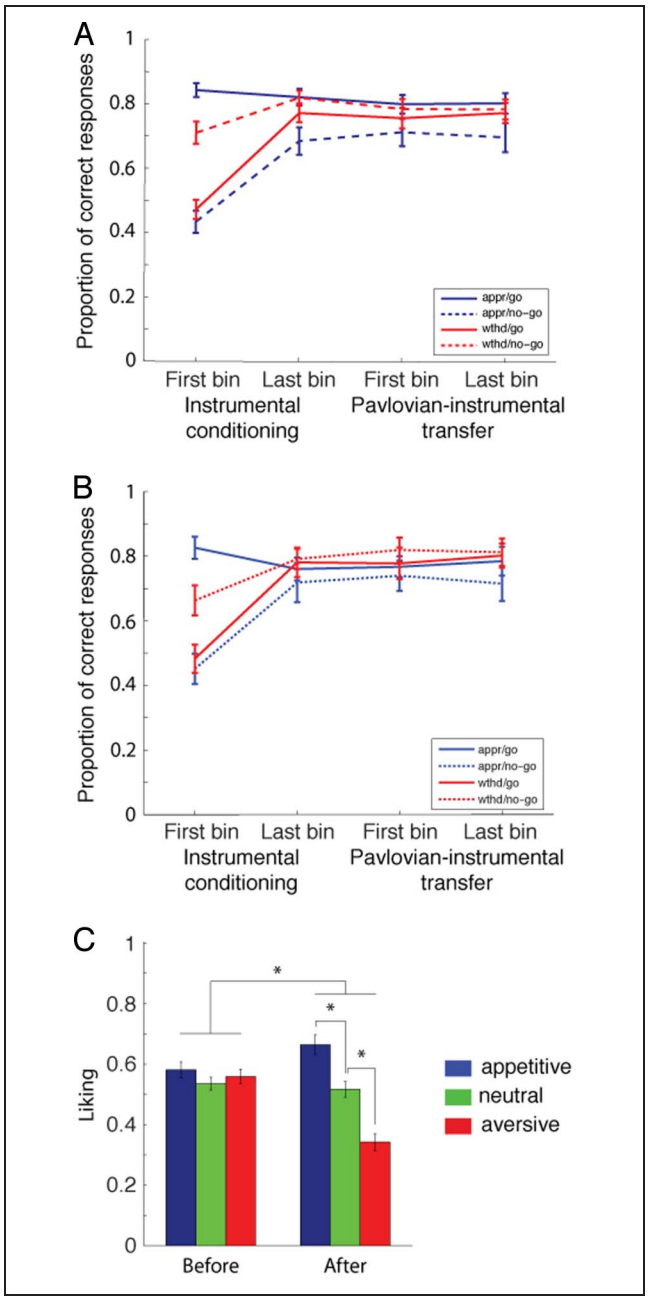

Figure 3. Instrumental learning and generalization to the PIT stage for (A) the whole group and for (B) the fMRI group separately. The proportion of correct choices are broken down by Response Type (go/no-go) and Action Context (approach/withdrawal). Error bars represent *SEM*s. (C) VAS ratings before and after Pavlovian conditioning. Bars represent group means of VAS scores (0 = *very aversive*, 0.5 = *neutral*, 1 = *very appetitive*). Error bars represent *SEM*s (*$p < .05$).

In addition, as is also explained by the interactions described in the previous paragraph, there was a significant interaction between Time Bin and Action Context, $F(1, 31) = 5.7, p < .05$, and a significant interaction between Response Type and Action Context across Time Bins, $F(1, 31) = 58.6, p < .001$. Furthermore, there was a main effect of Response Type, because of participants making more correct go responses than correct no-go responses across the instrumental training (main effect of Response Type: $F(1, 31) = 13.4, p < .01$). There was no significant main effect of or interaction with the factor Group.

For the fMRI group alone, almost the same pattern was found as for the whole group: Participants learned to make correct choices during the instrumental learning stage indicated by an increasing number of correct responses over time, $F(1, 17) = 50.0, p < .001$ (Figure 3B). Additionally there was a three-way interaction between Time Bin, Action Context, and Response Type, $F(1, 17) = 26.2, p < .001$. Again this was driven by the initial inclination of participants to collect the instrumental stimulus: Initially participants preferred to go during approach (paired sample t test: $t(17) = 5.3, p < .001$) but to no-go during withdrawal (paired sample t test: $t(17) = -2.4, p < .05$; simple interaction effect between Action Context and Response Type at Time Bin 1: $F(1, 17) = 35.6, p < .001$; Figure 3B). These biases were overcome at the end of the instrumental training stage (paired sample t test: $t(17) = 0.5, p > .1$; $t(17) = -0.2, p > .1$; simple interaction effect between Action Context × Response Type at Time Bin 2: $F(1, 17) = 0.2, p > .1$). In addition, there was a significant interaction between Response Type and Action Context across Time Bins, $F(1, 17) = 10.9, p < .01$.

Instrumental Generalization to the PIT Stage

Performance at the end of instrumental training generalized to and persisted throughout the PIT stage (Figure 3A): There were no significant main effects of or interactions with Time Bin when the two-level factor Time Bin was replaced with a Time Bin factor with three levels: the end of the instrumental training, the beginning of the PIT stage, and the end of the PIT stage. This was also the case when considering data from the fMRI group only (Figure 3B).

Pavlovian Conditioning

In both groups, analysis of the Pavlovian query trials confirmed successful Pavlovian conditioning (behavioral group: mean proportion correct = 95%, *SEM* = 3.1, range = 58–100%; fMRI group: mean proportion correct = 94%, *SEM* = 1.9, range = 80–100%).

A one-sample t test on the liking ratings of the US (i.e., juices; only available for the fMRI group) showed that, at baseline (pre), participants judged the aversive US to be aversive ($\text{mean}_{\text{pre}} = 0.21$, significantly different from 0.5: $t(17) = 14.4, p < .001$ [scores ranged from 0 (*aversive*) to 1 (*appetitive*) with 0.5 indicating *neutral*]) and the appetitive US to be appetitive (significantly different from 0.5: $\text{mean}_{\text{pre}} = 0.70, t(17) = 4.1, p = .001$). Ratings for the aversive US did not change significantly over the course of the experiment (pre vs. post, paired sample t test: $\text{mean}_{\text{post}} = 0.20, t(17) = 0.2, p > .05$); the appetitive US became slightly more appetitive across time (paired sample t test: $\text{mean}_{\text{post}} = 0.80, t(17) = 2.7, p < .05$; CS Valence × Time: $F(1, 17) = 5.1, p < .05$).

VAS ratings for the Pavlovian CSs (only available for the fMRI group) showed that Pavlovian conditioning induced changes in subjective liking (ANOVA Time × CS Valence: $F(1.3, 22.3) = 10.6, p = .002$; Figure 3C). Simple Time (pre/post) × CS Valence (two levels) interaction analyses confirmed that conditioning altered ratings for the aversive relative to the neutral CS, $F(1, 17) = 9.6, p = .007$, for the appetitive relative to the neutral CS, $F(1, 17) = 6.0, p = .026$, and for the appetitive relative to the aversive CS, $F(1, 17) = 12.4, p = .007$. There were no differences between the three CSs before conditioning (paired sample t test: appetitive vs. neutral, $t(17) = 1.5, p > .1$, appetitive versus aversive, $t(17) = 0.8, p > .1$, neutral versus aversive, $t(17) = -0.6, p > .1$). Conversely, after conditioning, liking ratings were significantly higher for the neutral than for the aversive CS, $F(1, 17) = 10.9, p < .01$, for the appetitive than for the neutral CS, $F(1, 17) = 9.7, p < .01$, and for the appetitive than for the aversive CS, $F(1, 17) = 24.5, p < .001$.

Pavlovian–Instrumental Transfer

Analysis of choice (go vs. no-go) data from the PIT stage revealed a significant action-specific PIT effect, which partially replicated that reported by Huys et al. (2011). Thus, the proportion of approach-go responses was lower during display of the aversive CS than that during display of the neutral CS (i.e., participants exhibited conditioned suppression). Conversely, the proportion of withdrawal-go responses was higher during display of the aversive CS than that during display of the neutral CS (Figure 4). This observation was confirmed statistically by a significant two-way interaction between Action Context (approach vs. withdrawal) and CS Valence (aversive vs. neutral; for the group as a whole: $F(1, 31) = 6.8, p < .05$; for the fMRI group only: $F(1, 17) = 3.3, p = .085$). Furthermore, simple effects analyses confirmed the presence of statistically significant simple effects of CS Valence (aversive vs. neutral) for approach (whole group: $F(1, 31) = 5.4, p < .05$; fMRI group only: $F(1, 17) = 2.1, p > .1$) as well as for withdrawal (whole group: $F(1, 31) = 5.1, p < .05$; fMRI group only: $F(1, 17) = 0.4, p > .1$). Thus, our task successfully revealed aversive PIT, an effect that was action specific.

In contrast, we did not find evidence for appetitive PIT. On the one hand, the omnibus F test with CS Valence as a three- instead of two-level factor (appetitive vs. neutral vs. aversive) did reveal a significant two-way interaction between Action Context and CS Valence (whole group:

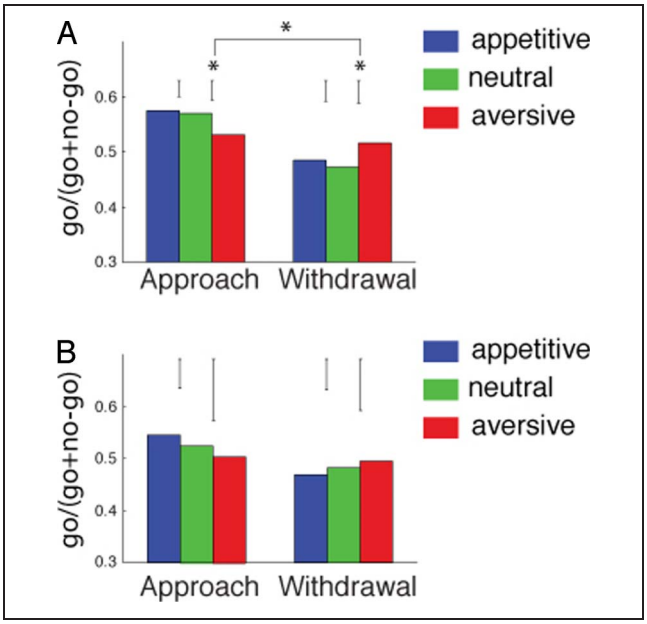

Figure 4. Behavioral data from the PIT stage. Shown are choice data as a function of Action Context (approach/withdrawal) and CS Valence (appetitive/neutral/aversive) for (A) the whole group and (B) the fMRI group separately. Error bars represent *SEM*s of the difference between, respectively, trials with appetitive and neutral CSs and trials with aversive and neutral CSs (*$p < .05$).

$F(2, 62) = 4.2, p < .05$; fMRI group only: $F(2, 34) = 2.3\, p > .1$ [linear contrast: $F(1, 17) = 3.7, p = .069$]), However, in contrast to our hypotheses, when appetitive CSs were compared with neutral CSs, there was no simple main effect of CS Valence (whole group: for approach: $F(1, 17) = 0.9, p > .1$; for withdrawal: $F(1, 17) = 0.4, p > .1$) and no simple interaction effect between Action Context and CS Valence (whole group: $F(1, 31) = 0.7, p > .1$). This suggests that our task was not appropriate for measuring appetitive PIT.

Irrespective of CS Valence, participants made more go-responses in the approach than in the withdrawal context (whole group: main effect of Action Context, $F(1, 31) = 4.4$, $p < .05$. This main effect of Action Context concurs with the pattern of performance in the initial instrumental training stage, which also revealed a main effect of Action Context.

There were no significant effects of the factor Group (behavioral/fMRI). Consistent with this lack of effect, the performance patterns were similar when analyzed separately for the fMRI group (Figure 4B), although the effects did not reach statistical significance (for stats see above).

There were no effects in terms of the total number of button presses (Table 1).

Imaging Data

BOLD Responses in the Amygdala and Nucleus Accumbens during Aversive PIT

We first performed an ANOVA using the parametric PIT regressors, and with Action Context (approach/withdrawal) and CS Valence (appetitive/neutral/aversive) as within-subject factors. There were no main effects of CS Valence and no interactions between Action Context and CS Valence, as revealed by whole-brain analyses and by small volume analyses (of the amygdala, nucleus accumbens, and the striatum).

However, when taking individual differences in behavioral PIT effects into account, we observed significant brain–behavior correlations in the amygdala and the nucleus accumbens: Participants who exhibited greater aversive inhibition of instrumental responding (across approach and withdrawal contexts) showed higher BOLD responses during aversive relative to neutral CSs (Figure 5). This was revealed by an ANOVA with the main regressors and the behavioral aversive PIT effect in terms of button presses as a covariate.

These brain–behavior correlations were because of significant associations between individual differences in the behavioral aversive PIT effect and BOLD responses in the bilateral amygdala and in the left nucleus accumbens. These effects in the amygdala and nucleus accumbens were present irrespective of Action Context. These analyses did not reveal any action-specific brain–behavior correlations, even when analyzed within our small volumes including the striatum. Thus, BOLD responses in the amygdala and nucleus accumbens to aversive CSs predicted individual differences in aversive Pavlovian inhibition, in a manner that was independent of Action Context.

The effects were also unique to aversive CSs and did not extend to appetitive CSs: A supplementary analysis contrasting appetitive and neutral CSs did not yield effects. Furthermore, supplementary analyses comparing aversive and appetitive CSs did not reveal the effects seen above in the comparison between aversive and neutral CSs. This lack of effect when including appetitive CSs might be because of increased variability during the appetitive CSs.

Ventromedial Prefrontal Cortex Differentiates between Approach and Withdrawal Context

Whole-brain ANOVA with the parametric PIT regressors (Action Context [approach/withdrawal] × CS Valence

Table 1. Average Number of Button Presses for the fMRI Group as a Function of Action Context (Approach/Withdrawal) and CS Valence (Appetitive/Neutral/Aversive) during the Pavlovian–Instrumental Transfer Stage (*SEM*)

	Action Context	
	Approach	*Withdrawal*
Appetitive	8.64 (0.27)	8.71 (0.30)
Neutral	8.78 (0.34)	8.66 (0.26)
Aversive	8.33 (0.32)	8.47 (0.36)

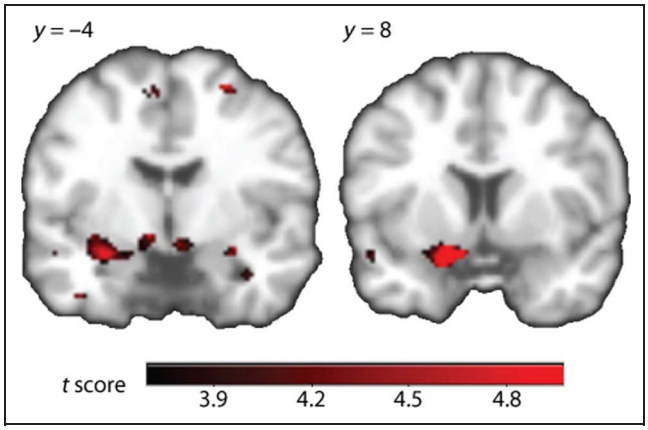

Figure 5. Aversive PIT-related BOLD response in the bilateral amygdala and left nucleus accumbens. The left image depicts regions of the amygdala (bilateral) where change in BOLD response between neutral and aversive CS trials was positively related to behavioral inhibition during aversive CS trials compared with neutral CS trials (small volume correction with the nucleus accumbens and amygdala VOI: $t = 5.45$, $p_{\text{FWE SV}} = .009$, MNI coordinates of peak voxel: $xyz = [-30\ -4\ -16]$; $t = 4.47$, $p_{\text{FWE SV}} = .044$, $xyz = [32\ -4\ -14]$, covariate: mean = 0.32, $SD = 0.71$). The right image shows that the same effect is significant for the left nucleus accumbens ($t = 4.97$, $p_{\text{FWE SV}} = .020$, $xyz = [-14\ 8\ -14]$). Images are displayed at a statistical threshold of $p < .001$ uncorrected.

[appetitive/neutral/aversive]) revealed a main effect of Action Context in the vmPFC (Figure 6). BOLD responses in this region were higher in the approach than in the withdrawal context. The inverse effect was observed in the bilateral lingual gyrus ($t = 9.57$, $p_{\text{FWE WB}} = .001$, MNI coordinates: $xyz = [16\ -74\ 0]$ and $[-2\ -78\ 18]$) and in the bilateral precuneus ($t = 6.0$, $p_{\text{FWE WB}} = .001$, $xyz = [10\ -54\ 48]$ and $[-10\ -48\ 48]$). Small volume analyses of responses in the amygdala, nucleus accumbens, and striatum did not reveal any subcortical action specificity.

Action Specificity of Aversive PIT Is Accompanied by Action-specific Fronto-striatal Connectivity

Next we assessed whether action specificity of behavioral aversive PIT was accompanied by PIT-related functional connectivity with this action context-specific BOLD response in the vmPFC. To this end, we conducted PPI analyses, separately for the approach and the withdrawal context, with the action-specific vmPFC region as the seed (Figure 6) and with a task contrast regressor representing aversive PIT (the number of button presses for aversive vs. neutral CSs).

When individual differences in behavioral PIT effects were not taken into account, small volume analyses revealed a significant effect in the striatum (centered on the caudate nucleus; MNI coordinates: $xyz = [-12\ 20\ 4]$) for withdrawal, but not approach. Specifically, in the withdrawal condition, there was a significant positive contribution of the vmPFC to the caudate nucleus during aversive PIT (Figure 7A). Thus, PIT-related connectivity between the vmPFC and the caudate nucleus was higher during aversive than during neutral CSs. No such effects, across the group as a whole, were found for the approach condition. However, when individual differences in behavioral PIT effects were taken into account, both small volume and even whole-brain analyses revealed a significant effect for the approach condition, again in the caudate nucleus (MNI coordinates: $xyz = [-4\ 12\ 12]$ and $[-14\ 26\ 2]$; Figure 7B). This effect reflected a negative association between the behavioral aversive PIT effect and the PPI effect: Greater aversive Pavlovian inhibition of approach responding was associated with reduced connectivity between the vmPFC and the caudate nucleus during aversive relative to neutral CSs. Thus, action-specific signal in the vmPFC contributed in a CS-dependent manner to the BOLD signal in the caudate nucleus. The same effect was significant in the bilateral nucleus accumbens (small volume correction with the nucleus accumbens and amygdala VOI: $t = 4.98$, $p_{\text{FWE SV}} = .017$, MNI coordinates: $xyz = [10\ 18\ -2]$; $t = 4.82$, $p_{\text{FWE SV}} = .017$, $xyz = [-12\ 10\ -8]$). This effect was unique to the striatum and the nucleus accumbens, as whole-brain and small volume correction analysis did not reveal any other meaningful effects.

DISCUSSION

This study addressed two key questions concerning human PIT. First, unlike prior studies, it revealed the neural mechanisms underlying PIT in the aversive domain and

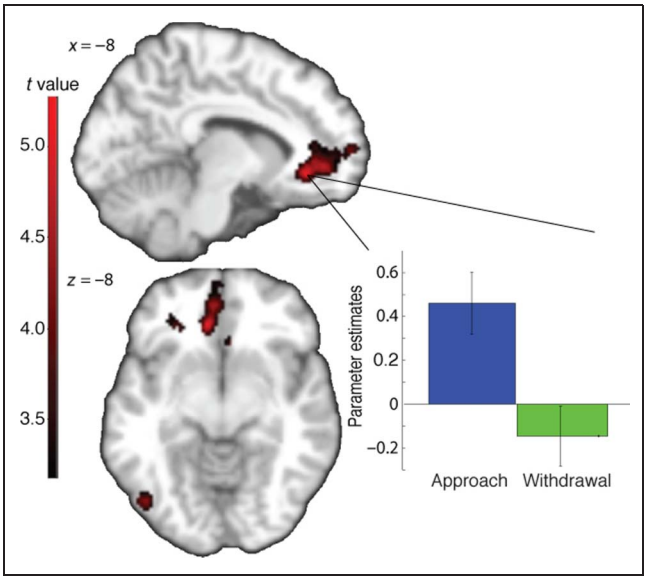

Figure 6. Action-specific BOLD response in the vmPFC. There was a main effect of Action Context in the vmPFC ($t = 5.25$, $p_{\text{FWE WB}} = .019$, MNI coordinates of peak voxel: $xyz = [-8\ 36\ -8]$). The bar graph shows parameter estimates from the peak voxel for the different Action Contexts. Images are displayed at a statistical threshold of $p < .001$ uncorrected.

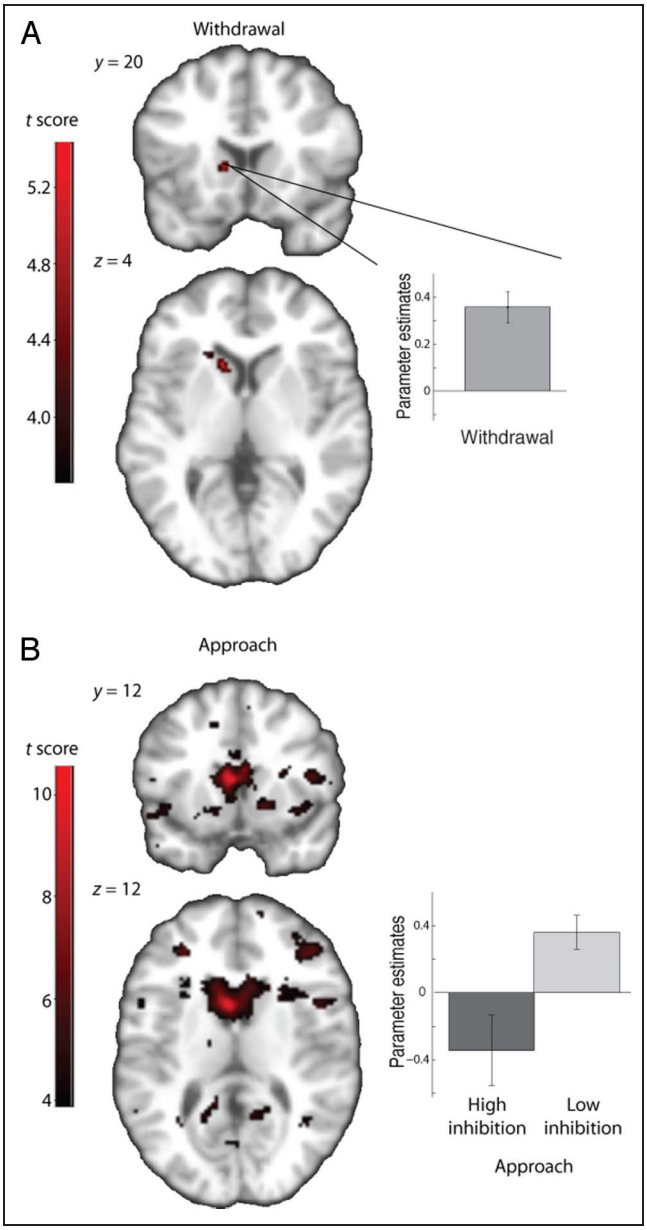

Figure 7. Functional connectivity during action-specific, aversive PIT. (A) The PPI analysis of aversive PIT in withdrawal showed PIT-related connectivity between the left caudate nucleus and the vmPFC (small volume correction with the striatum VOI: $t = 4.08$, $p_{\text{FWE SV}} = .031$, $xyz = [-12\ 20\ 4]$). The bar graph shows parameter estimates from the peak voxel. This reveals that PIT-related connectivity between the vmPFC and the caudate nucleus was higher during aversive than during neutral trials. (B) For aversive PIT in approach the brain image shows that PIT-related connectivity between the vmPFC and the caudate nucleus was associated with behavioral PIT effects (FWE correction for multiple comparisons for the whole brain: $t = 10.50$, $p_{\text{FWE WB}} = .001$, $xyz = [-4\ 12\ 12]$; $t = 9.75$, $p_{\text{FWE WB}} = .002$, $xyz = [-14\ 26\ 2]$, covariate: mean = 0.45, $SD = 1.05$). To interpret this association, parameter estimates from the peak voxel of the PPI analysis are shown in the bar graph for participants with high and low behavioral aversive PIT effects, that is, with high and low behavioral inhibition during presentation of the aversive CS (median split). This reveals that PIT-related connectivity between the vmPFC and the caudate nucleus was lower during aversive than during neutral trials for participants who showed more behavioral inhibition. Images are displayed at a statistical threshold of $p < .001$ uncorrected.

enabled us to conclude that the human amygdala and nucleus accumbens are involved in the effects of aversive Pavlovian cues on instrumental behavior. Second, this study addressed, for the first time, the neural mechanisms underlying action specificity of human PIT. Differential responses for approach and withdrawal were found in the vmPFC. Furthermore, aversive CSs modulated functional connectivity between the vmPFC and the caudate nucleus, both regions strongly associated with goal-directed instrumental control (Balleine & O'Doherty, 2010; Valentin, Dickinson, & O'Doherty, 2007). These results suggest that one origin of action specificity of PIT lies in the engagement of goal-directed control systems, such as the vmPFC and the caudate nucleus, and involves Pavlovian regulation of goal-directed fronto-striatal circuitry.

These findings generally concur with long established observations that the vmPFC is key for the affective control of behavior (Rushworth, Noonan, Boorman, Walton, & Behrens, 2011; Wallis, 2007; Clark & Manes, 2004; Greene, 2001; Damasio, 1997; Damasio & Everitt, 1996). Indeed, this region receives abundant input from regions that process affective information including the amygdala and the nucleus accumbens (Haber & Knutson, 2010; Haber, 2003; Ongür & Price, 2000; Mayberg et al., 1999), and it is critical for the instrumental guidance of behavior by representations of current goals (Valentin et al., 2007). Furthermore, recent electrophysiological findings in rats suggest that subsets of neurons in the vmPFC are involved in the integration of Pavlovian and instrumental information that underlies PIT (Homayoun & Moghaddam, 2009). This fMRI study did not reveal PIT signals in the vmPFC that evidence such integration. However, the pFC is well known not to act alone in guiding decision-making but interacts with a set of strongly connected subcortical structures via fronto-striatal circuits (Haber & Knutson, 2010; Haber, 2003; Alexander, DeLong, & Strick, 1986). In keeping with this, we found PIT-related connectivity between the vmPFC and the caudate nucleus as well as PIT-related signals in subcortical structures, such as the amygdala and nucleus accumbens.

Our study aimed specifically to address the neural mechanisms of action specificity in PIT. The finding that the vmPFC codes action specificity was obtained despite the fact that the values of approach and withdrawal goals (or actions) were the same (paired sample t test on action/Q values: $t(17) = -1.5$, $p > .1$). This is remarkable given previous work showing an important role for the vmPFC in representing goal (or action) values (Kahnt, Heinzle, Park, & Haynes, 2011; Hare, Camerer, Knoepfle, & Rangel, 2010; de Wit, Corlett, Aitken, Dickinson, & Fletcher, 2009; Hare, Camerer, & Rangel, 2009; Kable & Glimcher, 2009; Rangel, Camerer, & Montague, 2008). Its implication in goal-directed control is substantiated by another previous finding showing that BOLD responses in this region change as a function of outcome devaluation (Valentin et al., 2007). Our finding that approach behavior engages the vmPFC to a greater extent than does withdrawal

behavior might reflect the fact that, in this paradigm, there is an asymmetry between approach and withdrawal. Because the goal state (the instrumental stimulus) is more clearly delineated for approach than for withdrawal, it is conceivable that approach behavior is driven more readily by a goal-directed system (critically involving the vmPFC) than withdrawal behavior. According to an alternative, not mutually exclusive account, the differential response in the vmPFC might also reflect differences in visual attention paid to the goal state. Indeed, Lim, O'Doherty, and Rangel (2011) have recently shown that the vmPFC encodes (relative) value signals as a function of visual attention. This hypothesis also concurs with the finding that in our paradigm BOLD effects in visual occipital regions differentiated withdrawal from approach. Thus, action specificity in this PIT task might originate in systems that represent action values in a manner that is modulated by the goal state space and/or visual attention.

The observation that aversive PIT was accompanied by Pavlovian modulation of influences from this action-specific vmPFC on the caudate nucleus further strengthens the hypothesis that action specificity in PIT involves modulation of goal-directed control systems. Indeed, the rodent homologue of the caudate nucleus, that is, the dorsomedial striatum, has also been shown to be sensitive to changes in outcome devaluation (Yin, Knowlton, & Balleine, 2005; Yin, Ostlund, Knowlton, & Balleine, 2005). Furthermore, our findings reveal a strong relationship between the inhibition of instrumental approach by aversive Pavlovian cues and disruption of fronto-striatal connectivity by aversive Pavlovian cues. On the basis of this result, we speculate that aversive Pavlovian inhibition of approach (i.e., conditioned suppression) is accompanied by frontal suppression of striatal processing. The reverse pattern was observed for withdrawal, in which fronto-striatal connectivity was enhanced by the aversive cues, consistent with the speculation that aversive Pavlovian potentiation of withdrawal is accompanied by frontal enhancement of striatal processing. This proposal generally concurs with ideas that choice and planning of appropriate actions are instantiated by spiralling fronto-striatal pathways, including those connecting the vmPFC and the caudate nucleus (Balleine & O'Doherty, 2010; Haber, Fudge, & McFarland, 2000). Our connectivity findings indicate that processing in these pathways can be modulated by aversive Pavlovian CSs. This chimes well with our recent findings that inhibitory Pavlovian responses are able to significantly constrain goal-directed choice behavior (Huys et al., 2012).

The observation that the amygdala and the nucleus accumbens are involved in PIT concurs with animal studies showing that the influence of appetitive Pavlovian cues on instrumental decision-making depends on the integrity of the amygdala and nucleus accumbens (Corbit & Balleine, 2005, 2011). These studies have suggested that the amygdala represents the affective valence of Pavlovian cues, whereas the nucleus accumbens is thought to represent a limbic-motor interface, transmitting affective information to the spiraling cortico-striatal pathways. Our findings are also consistent with results from a study in humans revealing activity in both these regions during appetitive PIT (Talmi et al., 2008). That study showed that appetitive Pavlovian effects on instrumental vigour were associated with BOLD signal in the ventral striatum during appetitive cues compared with neutral cues. In addition, brain–behavior associations showed that participants who exhibited stronger behavioral PIT also exhibited stronger responses in the ventral striatum and amygdala. The key conclusion of this study is that these regions are also involved in aversive PIT. Unlike the pattern of responses in the vmPFC and unlike the pattern of connectivity with the caudate nucleus, the responses in the amygdala were not action specific, suggesting that it participates in Pavlovian inhibition of instrumental actions regardless of their approach/withdrawal nature.

The primary interest of this study was to uncover neural mechanisms of aversive rather than appetitive PIT. However, it is notable that, unlike prior work, this study did not replicate an effect of appetitive PIT in the amygdala or in the nucleus accumbens (cf. Talmi et al., 2008). We emphasize that our failure to demonstrate appetitive PIT does not diminish the validity of the paradigm for measuring aversive PIT. Nevertheless, in the following we consider a few hypotheses regarding this lack of effect. One key difference is that we used different outcomes for the Pavlovian and instrumental training stage and that our paradigm therefore captures exclusively outcome-general PIT. Talmi et al. (2008) used the same outcomes for both stages, and the effects they see in the nucleus accumbens and amygdala could therefore conceivably be driven by both outcome-general and outcome-selective PIT effects. Although animal work does suggest that both these regions are involved in outcome-general as well as outcome-selective PIT (Corbit & Balleine, 2005, 2011), the only extant study in humans on appetitive outcome-specific PIT did not find significant involvement of either the nucleus accumbens or amygdala (Bray et al., 2008). Thus, it may be that appetitive PIT BOLD signals in the human amygdala and accumbens are too weak to be observed in paradigms that tap into only outcome-specific or only outcome-general PIT. This could be addressed in future work by increasing the number of trials per subject. Another difference between our and previous work is that we used primary (i.e., appetitive juice) rather than secondary reinforcement (i.e., money) as Pavlovian USs. It is possible that involuntary reception of a juice while lying supine is not as appetitive as receiving money, although our subjective liking ratings did not suggest this was the case. Alternatively, extinction might have been faster for the appetitive than for the aversive juice.

Similar to Talmi et al. (2008), we found that PIT effects were more robust outside than inside the scanner. This replicated attenuation of PIT effects in, but not outside, the scanner, might reflect masking by non-specific factors. The scanner environment is loud and stressful and

may well mask subtle behavioral effects that depend on the display of background stimuli. Nevertheless, it should be noted that we did observe significant behavioral PIT over the group as a whole and, moreover, we also observed significant brain–behavior associations, strengthening our conclusion that the neural effects relate to behavioral PIT.

Our results suggest that outcome-general PIT involves affective regulation of goal-directed behavioral control systems. This generally concurs with the only PIT study in humans, which has shown that outcome-specific PIT can be sensitive to outcome-devaluation (Allman, DeLeon, Cataldo, Holland, & Johnson, 2010; see, however Holland, 2004, for different results in rodents). The current study suggests that, at least in humans, this might also hold for outcome-general PIT.

An understanding of how Pavlovian stimuli influence ongoing behavior may illuminate important aspects of pathological behavior. For example, one might conceptualize reactive aggression as seen in many mood (Monahan et al., 2001) or personality disorders (Coccaro, Sripada, Yanowitch, & Phan, 2011) as a potentation of aversive PIT. Aspects of proactive aggression, as seen in psychopathy (Cornell et al., 1996), might on the other hand reflect attenuated aversive PIT. This speaks to the notion that psychopathology could arise not only from abnormality within particular behavioral control systems, such as Pavlovian or goal-directed ones, but also from alterations in their interaction (Huys et al., 2012). Further exploration of these hypotheses will require experiments involving patient groups and precise characterization of interactions between the different behavioral control systems involved. As such, this study represents a stepping stone to future studies to advance our knowledge on affective, Pavlovian influences over instrumental behavior.

Acknowledgments

R. C. was supported by a VIDI grant from the Innovational Research Incentives Scheme of the Netherlands Organisation for Scientific Research (NWO).

Reprint requests should be sent to Dirk E. M. Geurts, Radboud University Nijmegen Medical Centre, Donders Institute for Brain, Cognition and Behavior, Centre for Cognitive Neuroimaging, Kapittelweg 29, 6500 HB, Nijmegen, The Netherlands, or via e-mail: d.geurts@donders.ru.nl.

REFERENCES

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience, 9,* 357–381.

Allman, M. J., DeLeon, I. G., Cataldo, M. F., Holland, P. C., & Johnson, A. W. (2010). Learning processes affecting human decision making: An assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes, 36,* 402–408.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35,* 48–69.

Bijttebier, P., Beck, I., Claes, L., & Vandereycken, W. (2009). Gray's reinforcement sensitivity theory as a framework for research on personality-psychopathology associations. *Clinical Psychology Review, 29,* 421–430.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Bray, S., Rangel, A., Shimojo, S., Balleine, B., & O'Doherty, J. P. (2008). The neural mechanisms underlying the influence of Pavlovian cues on human decision making. *Journal of Neuroscience, 28,* 5861–5866.

Büchel, C., Wise, R. J., Mummery, C. J., Poline, J. B., & Friston, K. J. (1996). Nonlinear regression in parametric activation studies. *Neuroimage, 4,* 60–66.

Clark, L., & Manes, F. (2004). Social and emotional decision-making following frontal lobe injury. *Neurocase, 10,* 398–403.

Coccaro, E. F., Sripada, C. S., Yanowitch, R. N., & Phan, K. L. (2011). Corticolimbic function in impulsive aggressive behavior. *Biological Psychiatry, 69,* 1153–1159.

Corbit, L. H., & Balleine, B. W. (2005). Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian–instrumental transfer. *Journal of Neuroscience, 25,* 962–970.

Corbit, L. H., & Balleine, B. W. (2011). The general and outcome-specific forms of Pavlovian–instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *Journal of Neuroscience, 31,* 11786–11794.

Cornell, D. G., Warren, J., Hawk, G., Stafford, E., Oram, G., & Pine, D. (1996). Psychopathy in instrumental and reactive violent offenders. *Journal of Consulting and Clinical Psychology, 64,* 783–790.

Damasio, A. R. (1997). Towards a neuropathology of emotion and mood. *Nature, 386,* 769–770.

Damasio, A. R., & Everitt, B. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]. *Philosophical Transactions of the Royal Society of London, 351,* 1413–1420.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711.

Dayan, P., & Huys, Q. J. M. (2008). Serotonin, inhibition, and negative mood. *PLoS Computational Biology, 4,* e4.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks: The Official Journal of the International Neural Network Society, 19,* 1153–1160.

de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A., & Fletcher, P. C. (2009). Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *Journal of Neuroscience, 29,* 11330–11338.

Di Giusto, J. A., Di Giusto, E. L., & King, M. G. (1974). Heart rate and muscle tension correlates of conditioned suppression in humans. *Journal of Experimental Psychology, 103,* 515–521.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255–278.

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., et al. (2011). A selective role for dopamine in stimulus-reward learning. *Nature, 469,* 53–57.

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution. *Neuroimage, 19,* 200–207.

Greene, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108.
Haber, S. N. (2003). The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy, 26,* 317–330.
Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *Journal of Neuroscience, 20,* 2369–2382.
Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology, 35,* 4–26.
Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience, 30,* 583–590.
Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324,* 646–648.
Holland, P. C. (2004). Relations between Pavlovian–instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes, 30,* 104–117.
Homayoun, H., & Moghaddam, B. (2009). Differential representation of Pavlovian–instrumental transfer by prefrontal cortex subregions and striatum. *European Journal of Neuroscience, 29,* 1461–1476.
Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., et al. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Computational Biology, 7,* e1002028.
Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology, 8,* e1002410.
Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: Consensus and controversy. *Neuron, 63,* 733–745.
Kahnt, T., Heinzle, J., Park, S. Q., & Haynes, J.-D. (2011). Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. *Neuroimage, 56,* 709–715.
Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *Journal of Neuroscience, 31,* 13214–13223.
Mayberg, H. S., Liotti, M., Brannan, S. K., McGinnis, S., Mahurin, R. K., Jerabek, P. A., et al. (1999). Reciprocal limbic-cortical function and negative mood: Converging PET findings in depression and normal sadness. *American Journal of Psychiatry, 156,* 675–682.
Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., et al. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York: Oxford University Press.
Ongür, D., & Price, J. L. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral Cortex (New York, N.Y.: 1991), 10,* 206–219.
Patenaude, B., Smith, S. M., Kennedy, D. N., & Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage, 56,* 907–922.
Poser, B. A., Versluis, M. J., Hoogduin, J. M., & Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. *Magnetic Resonance in Medicine, 55,* 1227–1235.
Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9,* 545–556.
Rushworth, M. F. S., Noonan, M. P., Boorman, E. D., Walton, M. E., & Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron, 70,* 1054–1069.
Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature, 450,* 102–105.
Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human Pavlovian–instrumental transfer. *Journal of Neuroscience, 28,* 360–368.
Trew, J. L. (2011). Exploring the roles of approach and avoidance in depression: An integrative model. *Clinical Psychology Review, 31,* 1156–1168.
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453–458.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage, 15,* 273–289.
Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience, 27,* 4019–4026.
Wallis, J. D. (2007). Neuronal mechanisms in prefrontal cortex underlying adaptive choice behavior. *Annals of the New York Academy of Sciences, 1121,* 447–460.
Weinstein, N. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820.
Worsley, K. J., & Friston, K. J. (1995). Analysis of fMRI time-series revisited—Again. *Neuroimage, 2,* 173–181.
Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *The European Journal of Neuroscience, 22,* 505–512.
Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *The European Journal of Neuroscience, 22,* 513–523.