The habenula encodes negative motivational value associated with primary punishment in humans

Rebecca P. Lawson[a,b,1], Ben Seymour[c,d], Eleanor Loh[b], Antoine Lutti[b,e], Raymond J. Dolan[b], Peter Dayan[f], Nikolaus Weiskopf[b], and Jonathan P. Roiser[a,1]

[a]Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, United Kingdom; [b]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom; [c]Computational and Biological Learning Laboratory, Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom; [d]Center for Information and Neural Networks, National Institute for Information and Communications Technology, Osaka 565-0871, Japan; [e]Laboratoire de Recherche en Neuroimagerie, Département des Neurosciences Cliniques, Centre Hospitalier Universitaire Vaudois, Université de Lausanne, 1011 Lausanne, Switzerland; and [f]Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, United Kingdom

Edited by John P. O'Doherty, California Institute of Technology, Pasadena, CA, and accepted by the Editorial Board June 24, 2014 (received for review December 18, 2013)

Learning what to approach, and what to avoid, involves assigning value to environmental cues that predict positive and negative events. Studies in animals indicate that the lateral habenula encodes the previously learned negative motivational value of stimuli. However, involvement of the habenula in dynamic trial-by-trial aversive learning has not been assessed, and the functional role of this structure in humans remains poorly characterized, in part, due to its small size. Using high-resolution functional neuroimaging and computational modeling of reinforcement learning, we demonstrate positive habenula responses to the dynamically changing values of cues signaling painful electric shocks, which predict behavioral suppression of responses to those cues across individuals. By contrast, negative habenula responses to monetary reward cue values predict behavioral invigoration. Our findings show that the habenula plays a key role in an online aversive learning system and in generating associated motivated behavior in humans.

high-resolution fMRI | conditioned behavior | pallidum

Learning which stimuli predict positive and negative outcomes, and thus should be approached or avoided, respectively, is central to an organism's ability to survive. Midbrain dopamine neurons respond to both unpredicted rewarding stimuli and to cues previously paired with rewards (1), consistent with behavioral approach toward those cues. As a counterpoint to these reward-related signals, neurons in the lateral habenula (LHb) of nonhuman primates respond to previously learned stimuli predicting the delivery of punishments and the omission of rewards, whereas they are inhibited by stimuli that signal upcoming rewards (2). These studies in nonhuman primates have concentrated on well-learned stimuli, and so have forsaken the opportunity to study the details of dynamic adaptation in the habenula. However, in many real-world scenarios, organisms learn about the motivational value of novel cues in their environment gradually, one exposure at a time, which raises the question as to whether the habenula plays a role in encoding the dynamically changing motivational value of cues that predict negative events.

Dynamic learning from aversive events permits the rapid experience-dependent updating of behavior, for example, the automatic suppression of approach, which is a characteristic of aversive conditioning (3). The LHb receives inputs from the globus pallidus (4), and its excitation inhibits midbrain dopamine neurons via the rostromedial tegmental nucleus (2). This position as a hub between corticolimbic networks and midbrain monoaminergic nuclei provides a means through which positively or negatively valenced stimuli can modulate motor output, leading to the hypothesis that the habenula plays a critical role in motivated behavior (5).

Studies using temporally precise optogenetic stimulation of the LHb in rodents provide convincing evidence that the habenula drives behavioral suppression (6). This structure has been suggested as a novel target for deep brain stimulation in the treatment of depression (7) based on the hypothesis that its overactivity might drive symptoms, such as disrupted decision making and anhedonia (8). Understanding the involvement of the habenula in generating negatively motivated behavior in humans is therefore central to our understanding of how the brain learns from and modifies behavior in response to aversive events, and its relevance for neuropsychiatric disorders, such as depression.

Investigating the habenula in humans with functional magnetic resonance imaging (fMRI) is nontrivial (9) due to its small size. Prior fMRI investigations have been limited by the use of standard data acquisition protocols, in which a single image volume element (volumetric pixel or voxel) is typically as large as the habenula itself. This low resolution, exacerbated by substantial spatial smoothing during standard data processing, is likely to induce localization error (9), rendering a signal from the habenula difficult to resolve from adjacent structures, such as the medial dorsal (MD) nucleus of the thalamus (10–12). Here, by using high-resolution fMRI, in conjunction with computational modeling of reinforcement learning in a paradigm that included

Significance

Organisms must learn adaptively about environmental cue–outcome associations to survive. Studies in nonhuman primates suggest that a small phylogenetically conserved brain structure, the habenula, encodes the values of cues previously paired with aversive outcomes. However, such a role for the habenula has never been demonstrated in humans. We establish that the habenula encodes associations with aversive outcomes in humans, specifically that it tracks the dynamically changing negative values of cues paired with painful electric shocks, consistent with a role in learning. Importantly, habenula responses predicted the extent to which individuals withdrew from or approached negative and positive cues, respectively. These results suggest that the habenula plays a central role in driving aversively motivated learning and behavior in humans.

Author contributions: R.P.L., B.S., R.J.D., P.D., and J.P.R. designed research; R.P.L. and E.L. performed research; R.P.L. and B.S. analyzed data; R.P.L., B.S., R.J.D., P.D., N.W., and J.P.R. wrote the paper; A.L. and N.W. developed imaging methods; and J.P.R. conceived the study.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.P.O. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: rebecca.lawson@ucl.ac.uk or j.roiser@ucl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1323586111/-/DCSupplemental.

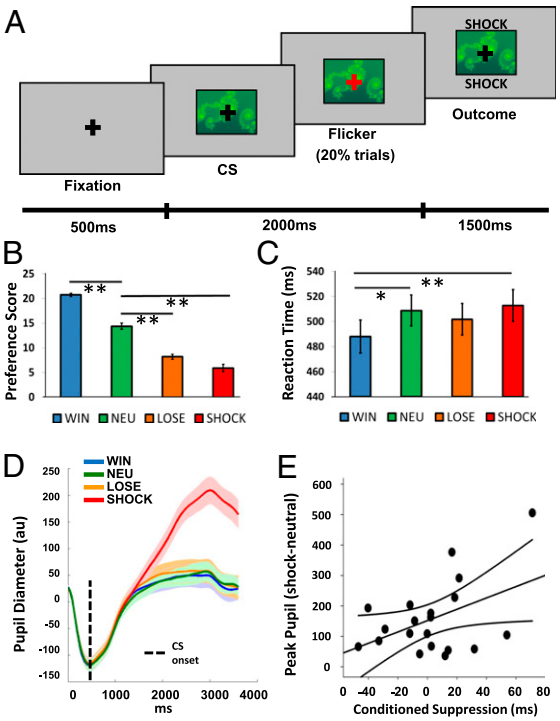

Fig. 1. Conditioning task and multiple indices of learning. (*A*) Exemplar trial (a detailed description is provided in the main text). (*B*) Explicit preference scores for win, loss, shock, and neutral (NEU) CSs (maximum score of 24). (*C*) Reaction times to respond to fixation flickers on win, loss, shock, and neutral CSs. (*D*) Pupil responses to win, loss, shock, and neutral CSs. (*E*) Relationship between autonomic (pupil responses to shock relative to neutral CSs) and implicit (conditioned suppression) measures of conditioning. Error bars and the shaded region in *D* represent SEMs. $*P < 0.01$; $**P < 0.005$.

primary punishments (painful electric shocks), we were able to test directly whether the habenula encodes changing associations with positive and negative stimuli over time in humans, and whether this encoding is linked to the modulation of behavioral output.

During fMRI, subjects ($n = 23$) performed a Pavlovian conditioning task in which they were passively exposed to seven abstract images [conditioned stimuli (CSs)] that were followed by different reinforcing outcomes (with high or low probability of reinforcement: win £1, lose £1, or painful electric shock, with the nonreinforced outcome being neutral, or a guaranteed neutral outcome) (Fig. 1*A* and *Materials and Methods*). During conditioning, subjects performed a fixation cross-flicker detection task to ensure attention (20% of trials, overlaid on CSs), which was independent of reinforcement. For the analysis of habenula responses, we used a model-based fMRI approach (13, 14), exploiting a reinforcement learning algorithm to calculate the trial-by-trial associative values of CSs that probabilistically predicted wins, losses, and shocks. We then used these values in the fMRI analysis as parametric regressors whose onsets were time-locked to the presentation of win, loss, and shock CSs. Because our central hypothesis related to the habenula, and given the small size and potential interindividual anatomical variability of this structure, we manually defined regions of interest (ROIs) on high-resolution anatomical scans for the left and right habenula in each subject according to a previously established protocol (9). This, and the use of high-resolution functional scans, enabled us to avoid signal contamination from adjacent structures, such as the MD thalamus. Additionally, our computational fMRI approach permitted us to investigate value-related responses in, and functional coupling with, regions that have known direct and indirect anatomical connections with the habenula, including the striatum and globus pallidus (4, 15). We therefore conducted additional exploratory whole-brain categorical and functional connectivity analyses enabling us to exploit our anatomically precise high-resolution data to examine how the habenula interacts with a wider network of brain regions known to play a crucial role in reinforcement learning.

Results

Behavioral Performance. We confirmed conditioning using three methods: explicitly (CS preference scores, measured after each block), implicitly (reaction times from the flicker detection task), and via autonomic responses [pupil dilation, measured using concurrent eye-tracking (16)]. Consistent with a pilot behavioral study (Fig. S1), all three approaches confirmed conditioning for shocks. Shock CSs were least preferred [significant effect of CS type: $F(1.87,41.30) = 97.28$, $P < 0.001$; Fig. 1*B*], were associated with slower responses [significant effect of CS type: $F(3,66) = 5.62$, $P = 0.002$; Fig. 1*C*], and elicited the largest peak pupil dilations [significant effect of CS type: $F(1.25,23.77) = 29.86$, $P < 0.001$; Fig. 1*D*]. Across subjects, the magnitude of pupil dilation to shock CSs (relative to neutral CSs) correlated positively with our behavioral measure of conditioned suppression [i.e., the slowing of responses on the flicker detection task during shock trials relative to neutral trials ($r = 0.45$, $P = 0.044$; Fig. 1*E*].

Habenula Responses to Negative CS Value. Analysis of blood oxygen level-dependent (BOLD) signals in the habenula, corresponding to computationally derived trial-by-trial fluctuations in CS values (Fig. 2*A*), revealed a significant linear effect of CS type [$F(1,22) = 4.34$, $P = 0.049$], which was qualified by a significant linear CS type * laterality interaction [$F(1,22) = 7.31$, $P = 0.013$]. Analysis of the right habenula only revealed a significant linear effect of CS type [$F(1,22) = 10.15$, $P = 0.004$], with planned pairwise comparisons showing that the response to parametrically varying shock CS values was significantly greater than to win CS values [$t(22) = 3.19$, $P = 0.004$], and also significantly different from zero [$t(22) = 2.35$, $P = 0.028$; Fig. 2*B*]. This latter result means that as

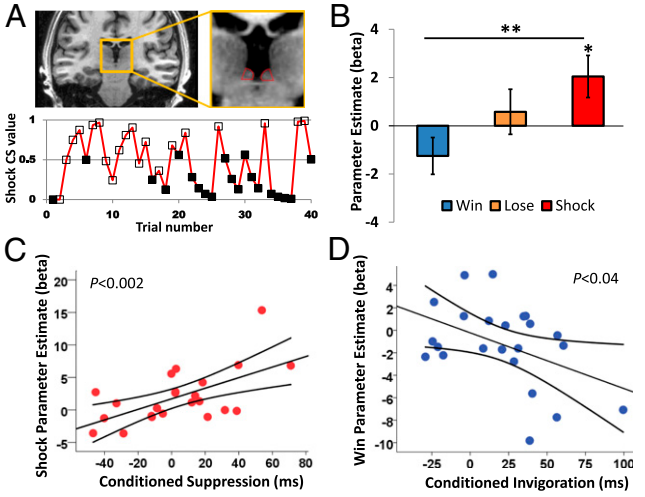

Fig. 2. Habenula results. (*A*) Location of the habenula on a coronal slice of a representative subject (*Upper*) and the trial-by-trial evolution of shock CS value during a single task block for a representative subject (*Lower*). Empty markers (□) indicate high-probability trials, and filled markers (■) indicate low-probability trials. (*B*) BOLD responses from the right habenula correspond to the dynamically changing values of win, loss, and shock CSs. (*C*) There is a positive correlation between the right habenula response to shock CS value and conditioned suppression. (*D*) There is a negative correlation between the right habenula response to win CS value and conditioned invigoration. Error bars represent SEM. $*P < 0.01$; $**P < 0.001$.

CSs become more predictive of shock, the response in the habenula increases linearly. Habenula responses to win and loss CS values were not significantly different from zero [win: $t(22) = -1.64$, $P = 0.12$; loss: $t(22) = 0.62$, $P = 0.54$]. Although the left habenula showed the same linear pattern of responses to win, loss, and shock CS values, the main effect of CS type was nonsignificant ($F < 1$; Fig. S2).

Relationship Between Habenula Responses and Behavior. If the habenula influences motor output (5), we would expect individual variability in our implicit conditioning measure to correlate with habenula responses, in a valence-specific manner. Strikingly, our behavioral measure of conditioned suppression was positively related to habenula responses to shock CS value across subjects [$r(23) = 0.60$, $P = 0.002$; Fig. 2*C*]. Furthermore, our behavioral measure of conditioned invigoration, the speeding of responses during the presentation of win CSs (relative to neutral CSs), was negatively related to habenula responses to win CS value [$r(23) = -0.44$, $P = 0.04$; Fig. 2*D*]. These correlations differed significantly from one another (Pearson–Filon $Z = 3.48$, $P < 0.001$).

MD Thalamus Responses. To determine whether signal from the MD thalamus, a comparatively large structure adjacent to the habenula, could be contributing to our effects, we drew left and right MD thalamus ROIs on the average normalized anatomical scan (*Materials and Methods*). BOLD responses to the computationally derived values of win, loss, and shock CSs were extracted in the same manner as for the habenula. We found no main effect of CS type ($F < 1$) and no interaction with laterality [$F(2,44) = 1.03$, $P = 0.37$; Fig. S3].

Habenula Responses to High- vs. Low-Probability Stimuli. To establish whether the habenula encodes a more general representation of (anti-) reward association, similar to that demonstrated in prior nonhuman primate studies (i.e., high vs. low probability of reinforcement, in addition to the trial-by-trial varying effects reported above), we ran another first-level model identical to that described above but with the addition of a second parametric modulator of the CS, which represented the contrast of the high- and low-probability CSs for each of the win, loss, and shock conditions.

We first confirmed that the results reported above for trial-by-trial fluctuations in CS value were unchanged in this analysis. Following a significant linear CS type * laterality interaction [$F(1,22) = 7.50$, $P = 0.012$], analysis of the right habenula only revealed a significant linear effect of CS type [$F(1,22) = 10.06$, $P = 0.004$], and planned comparisons confirmed that right habenula response to parametrically varying shock CS values was significantly greater than to win CS values [$t(22) = 3.17$, $P = 0.004$], and also significantly different from zero [$t(22) = 2.53$, $P = 0.019$]. The main effect of CS type was nonsignificant for the left habenula ($F < 1$).

We then examined habenula responses corresponding to the high- vs. low-probability categorical contrasts, which showed no interaction with laterality ($F < 1$). Collapsing across the left habenula and right habenula, we found a significant linear effect of CS type [$F(1,22) = 6.14$, $P = 0.021$]. The high- vs. low-probability contrasts for win and shock CSs were significantly different from each other [$t(22) = 2.48$, $P = 0.021$], and both showed a trend toward differing from zero (in opposite directions: win: $t(22) = -1.86$, $P = 0.08$; shock: $t(22) = 1.83$, $P = 0.08$; Fig. S4]. These results suggest that consistent with our finding that the habenula tracks trial-by-trial changes in shock CS value, the habenula may also encode a more general representation of negative value, similar to results reported previously in nonhuman primates (2).

Whole-Brain Analysis. To examine whether regions anatomically connected with the habenula also represent negative motivational value, we conducted a whole-brain analysis in normalized space. BOLD responses corresponding to computationally derived trial-by-trial shock CS values were detected in the vicinity of the medullary lamina of the left globus pallidus (peak voxel: [$x = -18$, $y = -6$, $z = 2$]; $Z = 3.10$, $P = 0.036$; small-volume corrected [SVC]; Fig. 3). Importantly, unlike most pallidal output, which is inhibitory, this region provides an excitatory input to the LHb in nonhuman primates (4, 17) and rats (18). Details of all other brain regions identified in this whole-brain analysis are presented in Table S1, and the corresponding negative contrasts can be found in Table S2.

Connectivity Analysis. To reveal how the habenula is functionally connected to other brain regions, we performed a psychophysiological interaction (PPI) analysis. We used the right habenula as the seed region for each subject, because the left habenula ROI showed no significant responses corresponding to CS value. Initially, we examined which brain regions are functionally connected to the habenula over the entire fMRI time series (i.e., separate from any CS value-dependent connectivity). At a whole-brain voxel-wise corrected significance level, we identified a large cluster showing positive connectivity, extending from the seed region to the left habenula, thalamus, and left pallidum ([$x = 7.5$, $y = -7.5$, $z = -3$], $Z = 5.20$). There was another large cluster in the right ventral striatum ([$x = 18$, $y = 12$, $z = -8$], $Z = 5.41$), extending to the bilateral medial wall of the caudate (left: [$x = -11$, $y = 5$, $z = 6$], $Z = 4.94$; right: [$x = 11$, $y = 6$, $z = 12$], $Z = 4.92$) and also the right amygdala ([$x = 27$, $y = 2$, $z = -12$], $Z = 4.80$; Fig. S5*A*). Several other regions survived whole-brain correction and are reported in Table S3.

The PPI analysis additionally allowed us to investigate patterns of functional coupling with the habenula as a function of changing CS value. At an exploratory threshold ($P < 0.005$ uncorrected, cluster size ≥10), we detected increased coupling with the right habenula as a function of increasing shock CS value in the left amygdala ([$x = -18$, $y = 2$, $z = -21$], $Z = 3.00$); bilateral posterior orbitofrontal cortex (pOFC) (left: [$x = -15$, $y = 15$, $z = -23$], $Z = 3.39$; right: [$x = 23$, $y = 12$, $z = -21$], $Z = 2.78$) and subcallosal anterior cingulate (Brodmann area [BA] 25: [$x = 3$, $y = 15$, $z = -12$], $Z = 3.29$; Fig. S5*B*). We provide these results (which did not fall within our a priori specified ROIs) for information only, without making inference, noting that they did not survive correction for multiple comparisons. Coupling with the right habenula increased as a function of increasing win CS value in the right ventral striatum extending to anterior putamen, which survived correction for multiple comparisons within our striatal ROI ([$x = 23$, $y = 18$, $z = -3$], $Z = 3.10$, $P = 0.028$ SVC; Fig. S5*C*). Other brain regions surviving the above exploratory threshold for both win and shock PPIs are presented in Table S3.

Discussion

Our results indicate that in humans, the habenula encodes the dynamically changing negative motivational value of stimuli that predict primary punishments. Importantly, these data go beyond prior findings in nonhuman primates (2, 19), which tested for

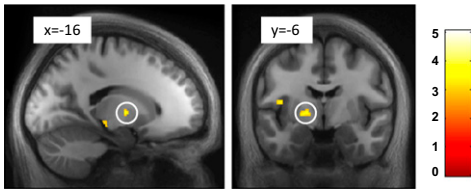

Fig. 3. Whole-brain analysis showing activation to shock CS value. Pallidal BOLD responses correspond to shock CS value. Images are thresholded at $P < 0.005$ (uncorrected) and at $k \geq 10$, and they are overlaid on the average normalized anatomical image; the color bar represents *t* values.

single-unit responses to previously overlearned stimuli. Our confirmatory analysis found some evidence of an additional correspondence with such responses in the habenula (Fig. S4); however, our computationally derived value regressors (Fig. 2*A*) show variation consistent with trial-by-trial learning that does not asymptote simply toward the true reinforcement probabilities. Consequently, our computational fMRI analysis uniquely demonstrates that the habenula represents the changing value of cues that predict reinforcers, as would be the case in naturalistic situations where organisms need to learn about dynamic cue–outcome associations from gradual exposure to environmental stimuli over time.

We overcame the limitations of standard fMRI acquisition for small subcortical brain structures by using high-resolution BOLD imaging in conjunction with anatomically precise ROIs, which were placed manually in native MRI space [on 770-μm anatomical images (9)]. This approach enables us to be confident that the signals we identified emanate from the habenula and not the neighboring MD thalamus. We confirmed this by showing that responses in the MD thalamus do not correspond to motivational value (Fig. S3). However, it is also worth noting that even with high-resolution fMRI, we do not have sufficient resolution to disambiguate the medial and lateral portions of the habenula, as outlined in our prior methodological paper on imaging the habenula in humans (9).

The linear response profile of the right habenula to increasingly aversive cues suggests that this region may provide a single mechanism for representing negative motivational value induced by both rewards and punishments. Although the habenula response to the value of win cues was significantly different from the value of shock cues, only the response to shock cues was significantly different from zero. It is possible that value coding by the habenula may be different for rewards and punishments; indeed, electrophysiological data in nonhuman primates support the notion that the representation of punishment in the LHb may be more precise than that of reward (19). Consistent with our results, the only previous high-resolution fMRI study of the habenula (which only examined the processing of appetitive stimuli) also failed to detect significant negative-going responses to the onset of reward-predicting cues (20).

It is also possible that the electric shocks, which were the most aversive outcomes in our study, framed the task such that nonshock cues were less motivationally salient, attenuating their associated neural responses, a suggestion supported by our pupil data, which showed greater dilation for shock-predictive cues relative to all other cues (Fig. 1*D*). Such contextual effects of primary and secondary reinforcers in aversive learning paradigms have previously been reported (21). We also note that, although still aversive, the average magnitude of shocks delivered in the scanning study was lower than in our behavioral pilot study (5.48 mA relative to 20.3 mA) to avoid discomfort-related movement, which would have corrupted our images. This may explain why there was no significant conditioned suppression at the group level in our scanning study, whereas there was in our behavioral pilot study (Fig. S1). The difference in reaction times between these conditions nonetheless reflects the extent to which a particular individual suppressed responses to the shock CS (relative to the neutral CS). It is this individual variability in reaction time that is predicted by habenula response to shock CS value, demonstrating a crucial link between habenula function and behavior.

Our main result demonstrates that the habenula tracks the values of cues that predict motivationally salient outcomes (primary punishments). A recent study demonstrated that inactivation of the LHb in rodents abolished subjective decision biases, effectively making choice behavior random (22). Our findings suggest that this effect could arise as the result of a failure to encode accurately the values of the available options during decision making, although further studies would be necessary to address this hypothesis. Furthermore, the finding that the right habenula alone showed robust responses to the value of shock cues is interesting in the wider context of laterality research on this structure in nonprimate species (23). The habenula shows phylogenetic conservation from fish to human and has attracted interest as a model for brain asymmetry, because many vertebrates show left–right differences in habenula size and neural circuitry (24). However, in our study, the electric shocks were always delivered to the left hand of subjects, and we speculate that this could provide a more parsimonious explanation of stronger responses to shock cue values in the contralateral LHb.

In addition to our primary analysis focused on the habenula, a whole-brain analysis revealed that globus pallidus responses also represent the value of shock cues. Interestingly, LHb-projecting neurons in this region are known to respond to punishment-predicting cues and nonreward-predicting cues in nonhuman primates, with pallidal responses occurring earlier than those in the LHb (4). Our data hint that LHb projecting pallidal neurons provide a driving input to this structure in humans, transmitting negative value-related information. Exploiting our high-resolution functional images and precisely placed anatomical ROIs, our connectivity analysis revealed that signal in the right habenula covaries with signal in a number of regions that have direct and indirect anatomical connections with the habenula (23). We found that the seed region, the right habenula, was strongly coupled with a large cluster extending into the left habenula. The left habenula and right habenula have a known direct connection, the habenular commissure, which likely mediates any contralateral functional connectivity. The finding that the habenula is functionally coupled with the pallidum is consistent with our whole-brain analysis of responses to shock CS value, as well as studies in rodents and nonhuman primates that have identified excitatory pallidal input to the LHb (4, 17, 18). Furthermore, we found that the habenula is functionally coupled with the striatum, including the medial wall of the caudate, which is strongly innervated by dopamine neurons (15) and has previously been implicated in fMRI studies of Pavlovian aversive learning (13).

Unfortunately, we did not have full coverage of the brainstem in our functional field of view (FOV); therefore, we were not able to investigate coupling between the habenula and midbrain dopaminergic nuclei. However, we did find the habenula to be functionally coupled with the amygdala, which has reciprocal connections with the substantia nigra (25). The substantia nigra is the main output of the LHb (5) and plays a crucial role in associative learning. One limitation of fMRI is that we are not able to infer whether the functional coupling detected with the habenula is inhibitory or excitatory. Nonetheless, these results provide evidence that the habenula operates within a network of brain regions known to participate in reinforcement learning (26).

In addition to the results discussed above, our PPI analysis provides very preliminary evidence that coupling between the habenula and the amygdala, pOFC, and BA25 increases as a function of shock CS value (Fig. S5*B*), consistent with the role of the latter regions in the acquisition of conditioned fear in both rodents and humans (27, 28). However, we note that these effects were detected at a liberal statistical threshold and did not survive stringent correction for multiple comparisons; we report them for completeness and they should be treated with caution until replicated. Furthermore, we found that coupling between the habenula and the striatum increased significantly as a function of win CS value (Fig. S5*C*), suggesting a role for habenula-striatal coupling in encoding information relating to reward value.

What is the functional role of value-related responses in the habenula? To answer this question, it is informative to consider how habenula responses relate to conditioned behavior. We identified a striking relationship across subjects between positive habenula responses to the value of shock cues and associated conditioned suppression and, conversely, between negative

habenula responses to the value of win cues and conditioned invigoration (Fig. 2 *C* and *D*). These data suggest that value-related responses in the habenula guide behavioral invigoration to rewards and suppression of behavior to punishments in humans, even when approach and withdrawal have no consequence. This accords with the view that the LHb output to the midbrain monoaminergic nuclei provides a critical pathway through which motor output can be modulated (5). This link with the invigoration and suppression of behavior hints at a potential role for the habenula in disorders characterized by aberrant motivated behavior, such as depression. Abnormalities in habenula structure and function have been reported in depressed patients (29, 30), as well as in animal models (31). Additionally, a recent study reported that glucose metabolism in the vicinity of the habenula decreased in depressed patients following treatment with ketamine (32). The data from the present study lend credence to the hypothesis that the habenula contributes to the generation of core depressive symptoms, especially those related to reinforcement processing, such as anhedonia and aberrant decision making (8).

Materials and Methods

Subjects. Twenty-seven subjects participated in this study. All had normal or corrected to normal vision, had no present or past neurological or psychiatric diagnosis, and provided written informed consent to participate. The study was approved by the London-Queen Square Research Ethics Committee, and subjects were compensated £50 for participation. Data were lost for two subjects due to scanner failure, and two subjects were removed from the analysis due to movement-induced image corruption, leaving 23 (15 female, mean age = 26 y, SD = 4.48, range = 20–37 y) participants in the analysis.

Experimental Procedures. ***Pain calibration.*** Pain was delivered to the left hand (fascia over adductor pollicis muscle) via a silver chloride electrode, using a single 1,000-Hz electrical pulse. Subjects underwent a thresholding procedure to control for heterogeneity in skin resistance and pain tolerance (33). Shocks were administered sequentially with step increases in amplitude, and subjects provided visual analog ratings of each shock on a scale from 0 (not painful) to 10 (terrible pain/pain that would cause me to move in the scanner). The level of shock delivered in the experiment was set to 80% of the maximum tolerated for each individual. The average shock strength was 5.48 (SD = 3.24) mA.

Conditioning paradigm. We used a Pavlovian paradigm with visual CSs (fractal images), probabilistically paired with win, loss, shock, or neutral outcome. There were seven CSs, associated with the following fixed outcome associations: 75% chance of £1 win, 25% chance of £1 win, 75% chance of £1 loss, 25% chance of £1 loss, 75% chance of shock, 25% chance of shock, and 100% no outcome (neutral). CSs were luminance-matched and assigned to conditions randomly across subjects. On trials where the reinforcing outcome (win, lose, or shock) was not presented, and on neutral trial outcomes, the word "nothing" was presented on-screen. The task is presented in Fig. 1*A*. On each trial, subjects initially saw a fixation cross, which remained on-screen for the entire trial; the CS appeared after 500 ms, remaining on-screen until the end of the trial; and the outcome was presented 2,000 ms following the CS onset. To ensure attention, on 20% of trials, the fixation cross present in the center of the screen flickered from black to red for 300 ms during CS presentation (but before outcome), and subjects were instructed to respond via a button press whenever this occurred. They were explicitly instructed that their responses made no difference to the outcomes they received. These trials were excluded from fMRI analysis. In total, 420 trials were presented over three blocks, which lasted 9.3 min each. Pilot reaction time data using this paradigm indicated robust conditioning (Fig. S1*A*).

Preference task. After each conditioning block, subjects' explicit knowledge of CS values was assessed using a preference task involving forced choices between pairs of CSs. Each CS was paired four times with every other CS, and subjects indicated which one they preferred. The position of each CS (on the left or right side of the screen) was randomized. The total number of preference choices for each CS was summed to calculate a total preference score (out of 24). Pilot data again indicated robust conditioning (Fig. S1*B*).

Pupillometry. Pupil diameter was measured during fMRI scanning by an IR eye tracker (Eyelink 1000; SR Research) recording at 500 Hz, and data were processed using custom-written algorithms in MATLAB R2011b (MathWorks). For each trial, blinks were treated with interpolation. Due to hardware failure, pupil data were not collected for one subject. Two subjects had more than one-third missing data on over one-third of trials and were removed from the analysis. For the remaining 20 subjects, we used the peak pupil response after presentation of the CS as a measure of autonomic arousal (16).

fMRI acquisition. MRI data were acquired with a 3-T Magnetom TIM Trio scanner (Siemens Healthcare) fitted with a 32-channel radio frequency receive head coil and body transmit coil. High-resolution, T2*-weighted, 2D echo-planar images (EPIs) were obtained using a custom-written sequence with the following parameters (34): matrix size of 128 × 128, FOV of 192 × 192 mm, in-plane resolution of 1.5 × 1.5 mm, interleaved slice order acquisition, slice thickness of 1.5 mm with no gap between slices, excitation flip angle of 90°, echo time (TE) of 36.2 ms, slice repetition time (TR) of 84.2 ms, and volume TR of 3.2 s. Thirty-eight slices were acquired with the FOV centered manually in line with the habenula in each subject. After reconstruction, three slices were discarded on either side of the encoding slab to avoid edge artifacts due to motion, leaving a total of 32 slices in each volume. Five dummy volumes were acquired before the image volumes to allow for T1 equilibration effects. Field maps were also acquired. Cardiac pulse signal and respiration were measured during EPI runs using a pulse oximeter and a pneumatic belt, respectively. These were used to correct for pulse- and respiration-related artifacts during analysis (see below) (35). High-resolution T1-weighted anatomical images were acquired using an optimized 3D modified driven equilibrium Fourier transform imaging sequence with correction for B1 inhomogeneities at 3 T (36). Image resolution was 770 μm isotropic (matrix size of 304 × 288 × 224, TR of 7.92 ms, TE of 2.48 ms, and excitation flip angle of 16°).

fMRI analysis. Statistical Parametric Mapping (SPM8; Wellcome Trust Centre for Neuroimaging, www.fil.ion.ucl.ac.uk/spm) was used to analyze all MRI data. For the ROI analysis of the habenula, each subject's data were slice time-corrected, realigned to the first image, unwarped using a field map of the static (B0) magnetic field (37), and coregistered to their individual anatomical scan, on which the habenula ROIs were placed according to a previously described procedure (9). Images were smoothed using a 2-mm FWHM Gaussian kernel to increase the signal-to-noise ratio without smoothing signal beyond the limits of the habenula ROI (9).

We used a reinforcement learning model to generate inferred values for the win, loss, and shock CSs on every trial (14). Specifically, we used a temporal difference model with a learning rate of $\alpha = 0.5$. This learning rate is supported by a number of studies examining both Pavlovian and instrumental learning (38, 39). Nonetheless, the results we acquired were robust to a range of learning rates (0.3–0.7; Fig. S6). In this model, the value (v) of a particular CS [referred to as a state (s)] is updated according to the following learning rule: $v(s + 1) \leftarrow v(s) + \alpha\delta$, where δ is the prediction error, defined as $\delta = r - v(s)$, and r is the outcome received.

At the subject level, fMRI data were analyzed in an event-related manner, using the general linear model, with the onsets of each win, loss, and shock CS (high- and low-probability stimuli combined in a single regressor) convolved with a synthetic hemodynamic response function in separate regressors. We used the model-based fMRI approach, in which the computationally derived CS values (see above and Fig. 2*A*) parametrically modulated the CS onset regressors on a trial-by-trial basis. In the model, we also included regressors for the onsets of win, loss, shock, and neutral outcomes, as well as realignment parameters to correct for subject movement and cardiac and respiration parameters to correct for physiological noise. A second model, in all other respects identical to the first, included a second parametric modulator of CS onset representing the contrast of high- vs. low-probability CS for each of the win, loss, and shock conditions. Note that our main inferences relate to the parametric regressors corresponding to the values of win, loss, and shock CSs, which are orthogonal to the regressors they modulate.

Group-level contrasts used the standard summary-statistics approach to random-effects analysis in SPM. Contrast estimates representing the win, loss, and shock CS values (i.e., the parametric modulator regressors from the subject level) were extracted from each individual's habenula ROI using the MarsBaR toolbox (40). Statistical tests conducted on these parametric contrast estimates at the group level indicate the reliability (across subjects) of the regression coefficient relating continuously varying CS value to habenula response at the subject level and, as such, do not require inclusion of a baseline condition because they already entail a contrast. For the exploratory whole-brain analysis, the respective contrast images for each subject were normalized to the standard space Montreal Neurological Institute template using the Dartel toolbox for SPM (41), smoothed with an 8-mm FWHM kernel, and included in group-level one-sample *t* tests thresholded at an exploratory threshold of $P < 0.005$ ($k \geq 10$). Small-volume correction was applied to a priori ROIs (described below).

Our PPI model included the deconvolved time series of signal in the right habenula ROI (physiological effect), a regressor corresponding to the parametric modulation of CS value at the time of CS onset (psychological effect),

and the product of these physiological and psychological regressors (the PPI) (42). We note that the psychological variable here is already an interaction between the parametric effect of CS value and the onset of the CS itself, because CS value is conditional upon CS onset and cannot strictly be isolated from it (because it is the expected value of a particular stimulus). For completeness, we also included regressors corresponding to the onsets of the CSs themselves in the PPI design matrix. Regressors were not orthogonalized before being entered into the design matrix. Separate PPI analyses were conducted for shock and win CS value regressors for each participant at the subject level. In addition to the realignment, cardiac, and respiration parameters, we included two nuisance time series: from a white-matter voxel in the center of the splenium of the corpus callosum and from a cerebrospinal fluid voxel in the center of the third ventricle occupying the same *y*-coordinate as the habenula. Contrast images corresponding to the main effect of the physiological variable and the PPI for win and shock CS values were normalized using the Dartel toolbox as described above and combined in group-level random-effects analyses. The former connectivity maps, representing the average linear effect of connectivity over all the levels of the psychological factor, were thresholded at $P < 0.05$ family-wise error corrected at the voxel level across the whole brain, whereas win and shock CS value PPI images were thresholded at an exploratory threshold of $P < 0.005$ ($k \geq 10$). Small-volume correction was applied to our a priori ROIs for the PPI analyses.

ROI definition. Habenula ROIs were placed manually for each subject in native space on high-resolution anatomical images according to a procedure previously described and validated (9). As a control region, the MD thalamus ROI was defined on the average normalized structural as a cylinder with a diameter of 4.5 mm that started on the same coronal slice as the habenula and continued anteriorly for 14 mm [approximately the length of the thalamus (43), including anterior and posterior MD thalamus regions], with the dorsolateral curve of the ROI following the dorsolateral edge of the MD thalamus against the third ventricle. ROIs applied to our whole-brain analyses for small-volume correction were the pallidum and ventral striatum. Our ventral striatum ROI was drawn as a sphere with a radius of 8 mm around a coordinate [$x = 20$, $y = 12$, $z = -8$] identified in a previous computational fMRI study of Pavlovian and instrumental learning (44), and the pallidum ROI was defined using a mask generated from the automated anatomical labeling atlas incorporated within the Wake Forest University PickAtlas toolbox for SPM (45).

Statistical analysis. Behavioral, peak pupil dilation, and habenula response data were analyzed in SPSS 20 (IBM). All data were inspected before analysis to check for deviations from Gaussian distributions. Differences between conditions were analyzed using repeated-measures ANOVA, and post hoc *t* tests (two-tailed). Where assumptions of heterogeneity of covariance were violated, degrees of freedom were corrected using the Greenhouse–Geisser approach. Correlations across subjects were assessed using Pearson's correlation coefficient (r), and differences in correlation coefficients were tested using the Pearson–Filon Z test (46).

ACKNOWLEDGMENTS. We thank David Bradbury, Alphonso Reid, and Oliver Josephs for help with the electric shock delivery and eye-tracking setup. We also thank Sanjay Manohar for assistance with pupillometry analysis and Ric Davis for data management. This research was supported by New Investigator Research Grant G0901275 from the Medical Research Council (to J.P.R.).

1. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593–1599.
2. Matsumoto M, Hikosaka O (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* 447(7148):1111–1115.
3. Estes WK, Skinner BF (1941) Some quantitative properties of anxiety. *J Exp Psychol* 29(5):390–400.
4. Hong S, Hikosaka O (2008) The globus pallidus sends reward-related signals to the lateral habenula. *Neuron* 60(4):720–729.
5. Hikosaka O (2010) The habenula: From stress evasion to value-based decision-making. *Nat Rev Neurosci* 11(7):503–513.
6. Stamatakis AM, Stuber GD (2012) Activation of lateral habenula inputs to the ventral midbrain promotes behavioral avoidance. *Nat Neurosci* 15(8):1105–1107.
7. Sartorius A, et al. (2010) Remission of major depression under deep brain stimulation of the lateral habenula in a therapy-refractory patient. *Biol Psychiatry* 67(2):e9–e11.
8. Sartorius A, Henn FA (2007) Deep brain stimulation of the lateral habenula in treatment resistant major depression. *Med Hypotheses* 69(6):1305–1308.
9. Lawson RP, Drevets WC, Roiser JP (2013) Defining the habenula in human neuroimaging studies. *Neuroimage* 64:722–727.
10. Ide JS, Li C-SR (2011) Error-related functional connectivity of the habenula in humans. *Front Hum Neurosci* 5:25.
11. Ullsperger M, von Cramon DY (2003) Error monitoring using external feedback: Specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. *J Neurosci* 23(10): 4308–4314.
12. Shelton L, et al. (2012) Mapping pain activation and connectivity of the human habenula. *J Neurophysiol* 107(10):2633–2648.
13. Seymour B, et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429(6992):664–667.
14. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38(2):329–337.
15. Haber SN, Knutson B (2010) The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology* 35(1):4–26.
16. Bitsios P, Szabadi E, Bradshaw CM (2004) The fear-inhibited light reflex: Importance of the anticipation of an aversive event. *Int J Psychophysiol* 52(1):87–95.
17. Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010) A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Neurophysiol* 104(2):1068–1072.
18. Shabel SJ, Proulx CD, Trias A, Murphy RT, Malinow R (2012) Input to the lateral habenula from the basal ganglia is excitatory, aversive, and suppressed by serotonin. *Neuron* 74(3):475–481.
19. Matsumoto M, Hikosaka O (2009) Representation of negative motivational value in the primate lateral habenula. *Nat Neurosci* 12(1):77–84.
20. Salas R, Baldwin P, de Biasi M, Montague PR (2010) BOLD Responses to Negative Reward Prediction Errors in Human Habenula. *Front Hum Neurosci* 4:36.
21. Delgado MR, Labouliere CD, Phelps EA (2006) Fear of losing money? Aversive conditioning with secondary reinforcers. *Soc Cogn Affect Neurosci* 1(3):250–259.
22. Stopper CM, Floresco SB (2014) What's better for me? Fundamental role for lateral habenula in promoting subjective decision biases. *Nat Neurosci* 17(1):33–35.
23. Bianco IH, Wilson SW (2009) The habenular nuclei: A conserved asymmetric relay station in the vertebrate brain. *Philos Trans R Soc Lond B Biol Sci* 364(1519): 1005–1020.
24. Amo R, et al. (2010) Identification of the zebrafish ventral habenula as a homolog of the mammalian lateral habenula. *J Neurosci* 30(4):1566–1574.
25. Lee HJ, et al. (2005) Role of amygdalo-nigral circuitry in conditioning of a visual stimulus paired with food. *J Neurosci* 25(15):3881–3888.
26. Garrison J, Erdeniz B, Done J (2013) Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neurosci Biobehav Rev* 37(7):1297–1310.
27. Milad MR, Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* 420(6911):70–74.
28. Phelps EA, Delgado MR, Nearing KI, LeDoux JE (2004) Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron* 43(6):897–905.
29. Roiser JP, et al. (2009) The effects of tryptophan depletion on neural responses to emotional words in remitted depression. *Biol Psychiatry* 66(5):441–450.
30. Savitz JB, et al. (2011) Habenula volume in bipolar disorder and major depressive disorder: A high-resolution magnetic resonance imaging study. *Biol Psychiatry* 69(4): 336–343.
31. Li B, et al. (2011) Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. *Nature* 470(7335):535–539.
32. Carlson PJ, et al. (2013) Neural correlates of rapid antidepressant response to ketamine in treatment-resistant unipolar depression: a preliminary positron emission tomography study. *Biol Psychiatry* 73(12):1213–1221.
33. Vlaev I, Seymour B, Dolan RJ, Chater N (2009) The price of pain and the value of suffering. *Psychol Sci* 20(3):309–317.
34. Lutti A, Thomas DL, Hutton C, Weiskopf N (2013) High-resolution functional MRI at 3 T: 3D/2D echo-planar imaging with optimized physiological noise correction. *Magn Reson Med* 69(6):1657–1664.
35. Hutton C, et al. (2011) The impact of physiological noise correction on fMRI at 7 T. *Neuroimage* 57(1):101–112.
36. Deichmann R, Schwarzbauer C, Turner R (2004) Optimisation of the 3D MDEFT sequence for anatomical brain imaging: Technical implications at 1.5 and 3 T. *Neuroimage* 21(2):757–767.
37. Hutton C, et al. (2002) Image distortion correction in fMRI: A quantitative evaluation. *Neuroimage* 16(1):217–240.
38. Seymour B, Daw N, Dayan P, Singer T, Dolan R (2007) Differential encoding of losses and gains in the human striatum. *J Neurosci* 27(18):4826–4831.
39. Seymour B, et al. (2005) Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat Neurosci* 8(9):1234–1240.
40. Brett M, Anton J-L, Valabregue R, Poline J-B (2002) Region of interest analysis using an SPM toolbox. 8th International Conference on Functional Mapping of the Human Brain. Available on CD-ROM in *NeuroImage* 16(2) (abstr).
41. Ashburner J (2007) A fast diffeomorphic image registration algorithm. *Neuroimage* 38(1):95–113.
42. Friston KJ, et al. (1997) Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* 6(3):218–229.
43. Mai J, Paxinos G, Voss T (2008) *Atlas of the Human Brain* (Academic, San Diego), 3rd Ed.
44. O'Doherty J, et al. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* 304(5669):452–454.
45. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19(3):1233–1239.
46. Raghunathan TE, Rosenthal R, Rubin DB (1996) Comparing correlated but nonoverlapping correlations. *Psychol Methods* 1(2):178–183.