Title: The habenula encodes motivation associated with primary punishment in humans

Authors: Rebecca P. Lawson[1, 2]*, Ben Seymour[2, 3], Eleanor Loh[2], Antoine Lutti[2,4], Raymond J. Dolan[2], Peter Dayan[5], Nikolaus Weiskopf[2] & Jonathan P. Roiser[1]*

Affiliations:

[1]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, WC1N 3AR, United Kingdom.

[2]Wellcome Trust Centre for Neuroimaging, University College London, 12 Queen Square, London, WC1N 3BG, United Kingdom.

[3]Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom.

[4]LREN, Département des neurosciences cliniques, CHUV, University of Lausanne, Lausanne, Switzerland.

[5]Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, WC1N 3AR, United Kingdom.

*Correspondence to: Rebecca Lawson, rebecca.lawson@ucl.ac.uk or Jonathan Roiser, j.roiser@ucl.ac.uk.

Abstract: 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 in humans the functional role of the habenula remains poorly characterized, in part due to its small size. Using high-resolution functional magnetic resonance imaging and computational modelling we demonstrate positive habenula responses to the dynamically changing value of cues signaling painful electric shocks, which predict behavioral suppression of responses to those cues across individuals. By contrast, habenula response to monetary reward cue value predicts behavioral invigoration. Our findings show that the habenula plays a key role in an online aversive learning system and in generating negatively motivated behavior in humans.

Statement of significance

Organisms must learn adaptively about environmental cue-outcome associations in order to survive. Studies in non-human 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 (painful electric shocks) in humans, and moreover that it tracks the dynamically changing negative values of cues, 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.

\body

Introduction

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, consistent with behavioral approach towards those cues(1). As a counterpoint to these reward-related signals, neurons in the lateral habenula (LHb) of non-human primates respond to previously learned stimuli predicting the delivery of punishments and the omission of primary rewards, while being inhibited by stimuli that signal upcoming rewards(2). These studies in non-human 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 more gradually, one exposure at a time. This 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 would permit 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, 5). Its position as a hub between corticolimbic 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 decision-making behavior(6).

Studies using temporally precise optogenetic stimulation of the LHb in rodents provide convincing evidence that the habenula drives behavioral suppression(7). This structure has been suggested as a novel target for deep brain stimulation in depression(8), based on an hypothesis that its over-activity might drive symptoms such as disrupted decision-making and anhedonia(9). Understanding the involvement of the habenula in generating negatively valenced 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 non-trivial(10). 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 of during standard data processing, is likely to induce localization error(10), rendering a signal from the habenula difficult to resolve from adjacent structures such as the medial dorsal (MD) nucleus of the thalamus(11–14). Here, by employing high-resolution fMRI(15), in conjunction with computational modeling of reinforcement learning in a paradigm that included primary punishments (painful electric shocks), we were able to test directly whether the habenula encodes changing negative motivational value across time in humans, and whether this encoding is coupled with the suppression or invigoration of behavioral output.

During fMRI, subjects (N=23) performed a Pavlovian conditioning task in which they were passively exposed to 7 abstract images (conditioned stimuli: CS) that were each followed by different reinforcing outcomes (with high or low probability reinforcement: win £1, lose £1, or painful electric shock, with the non-reinforced outcome being neutral; or a guaranteed neutral outcome) (**Fig. 1a, and Materials & 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 employed a model-based fMRI approach(1, 16, 17), 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. As our central hypothesis related to the habenula, and given the small size and potential inter-individual 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(10). 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, 18). 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 learning from rewards and punishments.

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 responses, measured using concurrent eye-tracking(19)). Consistent with a pilot behavioral study **(see 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. 1b),** were associated with slower responses (significant effect of CS type: F(3,66)=5.62, P=0.002, **Fig. 1c),** and elicited the largest peak pupil dilations (significant effect of CS type (F(1.25,23.77)=29.86, P<0.001, **Fig. 1d**). 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. 1e).**

Habenula responses to negative CS value

Analysis of blood-oxygen-level-dependent (BOLD) signals in the habenula, corresponding to computationally-derived trial-by-trial fluctuation in CS values **(Fig. 2a)**, revealed a significant linear effect of CS type (F(1,22)=4.34, P=0.049), and this was qualified by a significant linear CS type*laterality interaction (F(1,22)=7.31, P=0.013). Analysis of the right habenula revealed a significant linear effect of CS type (F(1,22)=10.15, P=0.004) with planned pairwise comparisons revealing the right habenula 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. 2b).** This means that as CSs become more shock-predicting, the response in the habenula increases**.** 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 non-significant (F<1, **Fig S2**).

Relationship between habenula responses and behavior

If the habenula influences motor output(6) then 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, namely slowing of responses during the presentation of shock CSs relative to neutral CSs, was positively related to habenula responses to shock CS value across subjects (r(23)=0.60, P<0.002**, Fig. 2c).** 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 2d**). These correlations differed significantly from one another (Pearson-Filon Z=3.48, P<0.001).

Medial dorsal thalamus responses

In order to determine whether signal from the medial dorsal (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 (see methods). BOLD responses to the computationally-defined 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 *versus* low probability stimuli

In order to establish whether the habenula encodes a more generic representation of (anti-) reward association (high vs. low probability of reinforcement), in addition to trial-by-trial varying effects, we ran another first-level model, identical to that described above, but with the addition of a second parametric modulator of the CS representing the contrast of high *versus* low probability for each of the win, lose and shock CSs. With this model we confirmed our previous results for-trial-by-trial fluctuations in CS value. Following a significant linear CS type*laterality interaction (F(1,22)=7.50, P=0.012), analysis of the right habenula 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 non-significant in the left habenula (F<1).

Habenula response to the high *versus* low probability contrast showed no interaction with laterality (F<1). Collapsing across left and right habenula revealed a significant linear effect of CS type (F(1,22)=6.14, P=0.021). The high *versus* low contrasts for win and shock CSs were significantly different to each other (t(22)=2.48, P=0.021) and both showed a trend towards 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 negative value, the habenula may also encode a more general representation of negative value, similar to results reported previously in non-human 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. Analysis of shock CS value revealed BOLD responses corresponding to computationally-derived trial-by-trial CS values in the vicinity of the medullary lamina of the left globus pallidus, which lies between the internal and external segments (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 non-human primates(4, 20) and rats(21). 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, as the left habenula ROI showed no 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 to any shock-value-dependent connectivity). At a whole-brain voxel-wise corrected significance level we identified a large cluster showing positive connectivity, comprising (as expected) the seed region itself, but extending 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. 4a)**. Several other regions survived whole-brain correction and are reported in **Table S3**.

The PPI analysis also allowed us to investigate patterns of functional coupling with the habenula as a function of changing CS value. At an exploratory threshold of P<0.005 uncorrected (k<10) we detected increased coupling with the right habenula as a function of 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 sub-callosal anterior cingulate (Broadmann area (BA) 25: [x=3 ,y=15, z=-12], Z=3.29; **Fig. S5)**. 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 win CS value in the right ventral striatum extending to anterior putamen, which survived correction for multiple comparisons ([x=23, y=18, z=-3], Z=3.10, P=0.028 SVC) (**Fig. 4b)**. Other brain regions surviving this exploratory threshold for both 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 findings in non-human primates (2, 22), which tested for single unit responses to previously over-learned stimuli. Confirmatory analysis found some evidence of an additional correspondence with such "tonic-like" responses in the habenula **(Fig. S3)**; however our computationally derived value regressors **(Fig. 2a)** show variation consistent with trial-by-trial learning that does not asymptote towards 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 employing 1.5 mm high-resolution BOLD imaging in conjunction with anatomically precise ROIs, which were placed manually in native MRI space (on 770 µm anatomical images(10)). 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 (see **Fig. S2**). 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 earlier paper on imaging the habenula in humans (10).

The linear response profile of the right habenula to increasingly aversive cues suggests that this region provides a single mechanism for representing negative motivational value induced by both rewards and punishments. While the habenula response to the value of win cues was significantly different to the value of shock cues, only the response to shock cues was significantly different from zero. This suggests that value coding by the habenula may be different for rewards and punishments, and indeed electrophysiological data in non-human primates support the notion that the representation of punishment in the LHb may be more precise than that of reward (22). Moreover, 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(23).

It is also possible that the electric shocks, which were the most aversive outcomes in our study, framed the task such that non-shock 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. 1d)**. Such contextual effects of primary and secondary reinforcers in aversive learning paradigms have previously been reported (24). We also note that, while still aversive, the average magnitude of shocks delivered in the scanning study was lower than in our behavioural pilot (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 behavioural pilot (Fig S1).

Our results demonstrate that, overall, the habenula is tracking the value of cues that predict salient outcomes (primary reinforcements). A recent study demonstrated that inactivation of the lateral habenula in rodents abolished subjective decision biases, effectively making choice behavior random (25). Our findings suggest that this effect could arise as the result of a failure to accurately encode the value of the available options during decision making, though further studies would be necessary to address this hypothesis. Furthermore, the finding that right habenula alone shows robust responses to the motivational value of shock cues is interesting in the wider context of laterality research on this structure in non-primate species (26). The habenula shows phylogenetic conservation from fish to human and has attracted interest as a model for brain asymmetry, since many vertebrates show left–right differences in habenula size and neural circuitry (27). 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 strong responses to shock cues in contralateral habenula.

In addition to our primary analysis focused on the habenula, a whole-brain analysis revealed that the globus pallidus also represents the value of shock cues. Interestingly LHb-projecting neurons in this region are known to respond to punishment predicting cues and non-reward predicting cues in non-human 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 with a number of regions that have direct and indirect anatomical connections with the habenula (26). We found that the seed region, the right habenula, was strongly coupled with itself as part of a large cluster extending into the left habenula. The left and right habenula have a known direct connection, the habenular commisure, which likely mediates any interhemispheric differences in functional connectivity. The finding that the habenula is coupled with the pallidum is consistent with our whole brain categorical analysis of responses to CS shock value, as well as studies in rodents and non-human primates that have identified excitatory pallidal input to the LHb(4, 20). 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(18) and has previously been implicated in fMRI studies of Pavlovian aversive learning(28).

Unfortunately, we did not have full coverage of the brainstem in our functional field of view and therefore 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 (29, 30), the main output of the lateral habenula ((6), that play a crucial role in associative learning. The limitation of fMRI is that we are not able to infer whether the functional coupling (in terms of BOLD signal) identified with the habenula is inhibitory or excitatory. Nonetheless, these results provide evidence that the habenula operates within a network of brain regions known participate in reinforcement learning(31).

In addition to the results discussed above, our PPI analysis provided provisional evidence suggesting that coupling between the habenula and the amygdala, pOFC and BA25 increases as a function of shock CS value (Fig. S5), consistent with the role of the latter regions in the acquisition of conditioned fear in both rodents and humans (32, 33)**.** However, we note that these effects were detected at a liberal statistical threshold and did not survive stringent correction for multiple comparisons: therefore 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. 4b),** 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? In order 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 response to the value of win cues and conditioned invigoration **(Fig. 2c, 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 lateral habenula output to the midbrain monoaminergic nuclei provides a critical pathway through which motor output can be modulated(6). This link with 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 (34, 35), as well as in animal models(36). Additionally, a recent study reported that glucose metabolism in the vicinity of the habenula decreased in depressed patients following treatment with ketamine(37). 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(9).

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 females, mean age 26 years, standard deviation (SD)=4.48, range=20-37) 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 1000 Hz electrical pulse. Subjects underwent a thresholding procedure to control for heterogeneity in skin resistance and pain tolerance (38). 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 mA (SD=3.24 mA).

Conditioning paradigm: We used a Pavlovian paradigm with visual CSs (fractal images), probabilistically paired with win, loss, shock or neutral outcomes. There were seven CSs, associated with the following fixed outcomes: 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; 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, the word “nothing” was presented on screen. The task is presented in **Fig. 1a**. 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 2000 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 functional magnetic resonance imaging (fMRI) analysis. In total 420 trials were presented over three blocks, which lasted 9.3 minutes each. Pilot behavioral data using this paradigm additionally indicated robust conditioning (**Fig. S1a**).

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, 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. S1b**).

Pupillometry: Pupil diameter was measured during fMRI scanning by an infrared eye tracker (SR Research Eyelink 1000) recording at 500 Hz, and data were processed using custom-written algorithms in Matlab R2011b (MathWorks, Natick, MA). 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 (19).

fMRI acquisition: MRI data were acquired with a 3T Magnetom TIM Trio scanner (Siemens Healthcare, Erlangen, Germany) 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(15): matrix size: 128x128; field-of-view (FOV): 192x192 mm; in-plane resolution: 1.5x1.5 mm; interleaved slice order acquisition; slice thickness: 1.5 mm with no gap between slices; excitation flip angle: 90°; echo time (TE): 36.2 ms; slice repetition time (TR): 84.2 ms; volume TR 3.2s. 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 prior to 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) (39). High-resolution T1-weighted anatomical images were acquired using an optimized 3D MDEFT imaging sequence with correction for B1 in-homogeneities at 3T(40). Image resolution was 770μm isotropic (matrix size: 304 × 288 × 224; TR: 7.92 ms; TE: 2.48 ms; excitation flip angle: 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 magnetic field (B0) (41) and co-registered to their individual structural scan, on which the habenula ROIs were placed according to a previously described procedure (10). Images were smoothed using a Gaussian kernel with full width at half maximum (FWHM) of 2 mm to increase signal-to-noise without smoothing signal beyond the limits of the habenula ROI (10).

We used a reinforcement learning model to generate inferred values for the win, loss and shock CSs on every trial(16). Specifically, we used a temporal difference model with a learning rate of α=0.5. This learning rate is supported by a number of studies, examining both Pavlovian and instrumental learning(17, 42). Nonetheless, the results we acquired were robust to a range of learning rates (0.3–0.7, see **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 CSs (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 paragraph) parametrically modulated the CS onset regressors on a trial-by-trial basis. We also included in the model 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, identical to this one, included a second parametric modulator of CS onset representing the contrast high *versus* low probability for each of the win, lose and shock cues. 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). Summary statistics conducted on these contrast estimates indicate the statistical reliability of the regression coefficient relating continuously varying CS value to habenula response and, as such, do not necessitate a baseline comparison. For the exploratory whole-brain analysis the respective contrast images for each subject were normalized to the standard space Montreal Neurological Institute (MNI) template using the Dartel toolbox for SPM(43), smoothed with an 8mm full width at half maximum kernel and included in group-level one-sample t-tests thresholded at an exploratory threshold of $P<0.005$ (k=10). SVC was applied to *a priori* regions of interest as described in the ROI definition section below.

We performed a psychophysiological interaction (PPI) analysis by extracting the deconvolved time series of signal in the right habenula ROI (physiological effect), a regressor corresponding to the parametric modulation of CS value (psychological effect), and taking the product of these physiological and psychological regressors (the psychophysiological interaction)(44). The psychological variable is already an interaction between the parametric effect of CS value and the onset of the CS itself since CS value is conditional upon cue onset and cannot strictly be isolated from it (as it is the expected value of a particular CS). The PPI design matrix also included nuisanse regressors corresponding to the onsets of the CSs themselves**.** Separate PPI analyses were conducted for shock and win CS value regressors for each participant at the subject level. In addition to the movement and physiological (cardiac and respiration) parameters we also 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 value 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, while win and shock CS value PPI images were thresholded at an exploratory threshold of $P<0.005$ (k=10). Small volume correction (SVC) was applied to our *a priori* regions of interest for the PPI analyses.

ROI definition: Habenula ROIs were placed manually for each subject in native space on high-resolution structural images according to a procedure previously described and validated(10). 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(45) including anterior and posterior MD thalamus regions), with the top-left curve of the ROI following the dorsal edge of the MD thalamus against the third ventricle. ROIs applied to our whole-brain analysis were the pallidum and ventral striatum. Our ventral striatum ROI was drawn as a sphere of 8mm radius around a coordinate [x=20, y=12, z=-8] identified in a previous computational-fMRI study of Pavlovian and Instrumental learning(46) and the pallidum ROI was defined using a mask generated from the automated anatomical labeling atlas incorporated within the WFU PickAtlas toolbox for SPM(47).

Statistical analysis: Behavioral, peak pupil dilation and habenula response data were analyzed in SPSS 20 (IBM, Chicago, IL). All data were inspected prior to analysis to check for deviations from Gaussian distributions. Differences between conditions were analyzed using repeated-measures analysis of variance (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 r, and differences in correlation coefficients were tested using the Pearson-Filon Z test(48).

Acknowledgments: This research was supported by a New Investigator Research Grant from the Medical Research Council (G0901275) to JPR. We thank David Bradbury, Alphonso Reid and Oliver Josephs for help with the electric shock delivery and eye tracking set up and Sanjay Manohar for assistance with pupillometry analysis.

References:

1. Schultz W, Dayan P, Montague PR (1997) A Neural Substrate of Prediction and Reward. *Science* 275:1593–1599.

2. Matsumoto M, Hikosaka O (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* 447:1111–1115.

3. Estes WK, Skinner BF (1941) Some quantitative properties of anxiety. *Journal of Experimental Psychology* 29:390–400.

4. Hong S, Hikosaka O (2008) The Globus Pallidus Sends Reward-Related Signals to the Lateral Habenula. *Neuron* 60:720–729.

5. Jhou TC, Geisler S, Marinelli M, Degarmo BA, Zahm DS (2009) The mesopontine rostromedial tegmental nucleus: A structure targeted by the lateral habenula that projects to the ventral tegmental area of Tsai and substantia nigra compacta. *J Comp Neurol* 513:566–596.

6. Hikosaka O (2010) The habenula: from stress evasion to value-based decision-making. *Nat Rev Neurosci* 11:503–513.

7. Stamatakis AM, Stuber GD (2012) Activation of lateral habenula inputs to the ventral midbrain promotes behavioral avoidance. *Nat Neurosci* 15:1105–1107.

8. Sartorius A et al. (2010) Remission of Major Depression Under Deep Brain Stimulation of the Lateral Habenula in a Therapy-Refractory Patient. *Biological Psychiatry* 67:e9–e11.

9. Sartorius A, Henn FA (2007) Deep brain stimulation of the lateral habenula in treatment resistant major depression. *Medical Hypotheses* 69:1305–1308.

10. Lawson RP, Drevets WC, Roiser JP (2013) Defining the habenula in human neuroimaging studies. *NeuroImage* 64:722–727.

11. Noonan MP, Mars RB, Rushworth MFS (2011) Distinct Roles of Three Frontal Cortical Areas in Reward-Guided Behavior. *The Journal of Neuroscience* 31:14399 –14412.

12. Ide JS, Li C-SR (2011) Error-Related Functional Connectivity of the Habenula in Humans. *Front Hum Neurosci* 5. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3060701/ [Accessed June 22, 2011].

13. 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. *The Journal of Neuroscience* 23:4308 –4314.

14. Shelton L et al. (2012) Mapping pain activation and connectivity of the human habenula. *Journal of Neurophysiology* 107:2633–2648.

15. Lutti A, Thomas DL, Hutton C, Weiskopf N (2012) High-resolution functional MRI at 3 T: 3D/2D echo-planar imaging with optimized physiological noise correction. *Magn Reson Med*:n/a.

16. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron* 38:329–337.

17. Seymour B, Daw N, Dayan P, Singer T, Dolan R (2007) Differential Encoding of Losses and Gains in the Human Striatum. *The Journal of Neuroscience* 27:4826–4831.

18. Haber SN, Knutson B (2009) The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology* 35:4–26.

19. Bitsios P, Szabadi E, Bradshaw C. (2004) The fear-inhibited light reflex: importance of the anticipation of an aversive event. *International Journal of Psychophysiology* 52:87–95.

20. Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010) A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*. Available at: http://jn.physiology.org/content/early/2010/06/10/jn.00158.2010.abstract.

21. 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* 74:475–481.

22. Matsumoto M, Hikosaka O (2009) Representation of negative motivational value in the primate lateral habenula. *Nature Neuroscience* 12:77–84.

23. Salas R, Baldwin P, De Biasi M, Montague R (2010) BOLD responses to negative reward prediction errors in the human habenula. *Front Hum Neurosci* 36.

24. Delgado MR, Labouliere CD, Phelps EA (2006) Fear of losing money? Aversive conditioning with secondary reinforcers. *Social cognitive and affective neuroscience* 1:250–259.

25. Stopper CM, Floresco SB (2014) What's better for me[quest] Fundamental role for lateral habenula in promoting subjective decision biases. *Nat Neurosci* 17:33–35.

26. Bianco IH, Wilson SW (2009) The habenular nuclei: a conserved asymmetric relay station in the vertebrate brain. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364:1005–1020.

27. Amo R et al. (2010) Identification of the Zebrafish Ventral Habenula As a Homolog of the Mammalian Lateral Habenula. *The Journal of Neuroscience* 30:1566–1574.

28. Seymour B et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429:664–667.

29. Lee HJ et al. (2005) Role of amygdalo-nigral circuitry in conditioning of a visual stimulus paired with food. *The Journal of neuroscience* 25:3881–3888.

30. Lee HJ, Youn JM, Gallagher M, Holland PC (2006) Role of substantia nigra–amygdala connections in surprise-induced enhancement of attention. *The Journal of neuroscience* 26:6077–6081.

31. Garrison J, Erdeniz B, Done J (2013) Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews* 37:1297–1310.

32. Milad MR, Quirk GJ (2002) Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* 420:70–74.

33. Phelps EA, Delgado MR, Nearing KI, LeDoux JE (2004) Extinction Learning in Humans: Role of the Amygdala and vmPFC. *Neuron* 43:897–905.

34. Roiser JP et al. (2009) The Effects of Tryptophan Depletion on Neural Responses to Emotional Words in Remitted Depression. *Biological Psychiatry* 66:441–450.

35. Savitz JB et al. (2011) Habenula volume in bipolar disorder and major depressive disorder: a high-resolution magnetic resonance imaging study. *BIOL PSYCHIATRY* 69:336–343.

36. Li B et al. (2011) Synaptic potentiation onto habenula neurons in the learned helplessness model of depression. *Nature* 470:535–539.

37. Carlson PJ et al. (2013) Neural Correlates of Rapid Antidepressant Response to Ketamine in Treatment-Resistant Unipolar Depression: A Preliminary Positron Emission Tomography Study. *Biological Psychiatry* 73:1213–1221.

38. Vlaev I, Seymour B, Dolan RJ, Chater N (2009) The Price of Pain and the Value of Suffering. *Psychological Science (Wiley-Blackwell)* 20:309–317.

39. Hutton C et al. (2011) The impact of physiological noise correction on fMRI at 7 T. *NeuroImage* 57:101–112.

40. 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* 21:757–767.

41. Hutton C et al. (2002) Image Distortion Correction in fMRI: A Quantitative Evaluation. *NeuroImage* 16:217–240.

42. Seymour B et al. (2005) Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nat Neurosci* 8:1234–1240.

43. Ashburner J (2007) A fast diffeomorphic image registration algorithm. *NeuroImage* 38:95–113.

44. Friston K. et al. (1997) Psychophysiological and Modulatory Interactions in Neuroimaging. *NeuroImage* 6:218–229.

45. Mai J., Paxinos G, Voss T (2008) *Atlas of the Human Brain* (Academic Press: Elsevier). 3rd Ed.

46. O'Doherty J et al. (2004) Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science* 304:452–454.

47. Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage* 19:1233–1239.

48. Steiger JH (1980) Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87:245–251.

Figure legends

Fig. 1: *Conditioning task and multiple indices of learning.* (**a**) An exemplar trial. Subjects were shown abstract images (CSs) predictive of different outcomes in a probabilistic manner (shock outcome shown here, see text for details). The fixation cross was always present and subjects were instructed to respond via a button press whenever it briefly (300ms) flickered from black to red (20% of trials). The fixation cross never flickered during the outcome phase and subjects were explicitly instructed that the flickers and their responses bore no relation to the outcomes they would receive; **(b)** explicit preference scores for win, loss, shock and neutral CSs (maximum score 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)** the relationship between autonomic (pupil responses to shock relative to neutral CSs) and our implicit measure

of aversive conditioning (reaction times during shock CSs relative to neutral CSs). Error bars, and shaded region in (d), represent standard errors of the mean (SEM). * P<0.01, ** P<0.005.

Fig. 2: *Habenula region of interest results.* **(a)** Location of the habenula on a coronal slice of a representative subject (top); and the trial-by-trial evolution of shock CS value during a single task block for a representative subject (bottom). Empty markers indicate high-probability trials and filled black markers indicate low-probability trials; **(b)** extracted BOLD responses from the right habenula corresponding to the dynamically changing value of win, loss and shock CSs, averaged across subjects; **(c)** positive correlation between right habenula response to shock CS value and conditioned suppression – the lengthening of reaction times when responding to flickers superimposed on shock CSs (relative to neutral CSs); **(d)** negative correlation between right habenula response to win CS value and conditioned invigoration – the shortening of reaction times when responding to flickers superimposed on shock CSs (relative to neutral CSs). Error bars represent SEM. * P<0.01; ** P<0.001.

Fig. 3: *Whole-brain fMRI analysis showing activation to the value of shock CSs.* Pallidal BOLD responses corresponding to shock CS value (white circle). Image thresholded at P<0.005 (uncorrected) and overlaid on the average normalized structural image; the color bar represents t-values.

Fig. 4: *PPI analysis results*. **(a)** 'seed based' connectivity (main effect of the physiological variable) over the entire fMRI time series between the right habenula (seed region) and the right ventral striatum, bilateral medial wall of the caudate and the globus pallidus [images thresholded at P<0.05, whole-brain Family-Wise Error corrected at the voxel-level]; **(b)** increased coupling between the right habenula and the right ventral striatum as a function of increasing win CS value. Image thresholded at exploratory P < 0.005, cluster threshold (k) > 10]. The striatum was included in our a priori ROUs and this result survived small volume correction. All images overlaid on the average normalized structural image; color bars represent t-values.

Fig. S1: *Pilot data – indices of conditioning*. Pilot data from a study conducted outside the scanner (N=20; 10 female; mean age=38.5 (SD=8.03). Subjects underwent a thresholding procedure as in the main study, rating each shock on a visual analogue scale from *0 – not painful*, to *10 – worst imaginable pain.* As subject movement was not a consideration in this pilot study the average shock strength tolerated by subjects was higher, 20.3 mA (SEM=4.76 mA). **(a)** Subjects responded slowest in the flicker detection task when shock CSs were on screen (significant effect of CS type: F(3,57)=4.58, P=0.006); **(b)** subjects preferred to see loss and shock CSs least (significant effect of CS type: F(3,57)=105.56, P<0.001). Error bars represent SEM. * P<0.01; ** P<0.001.

Fig. S2: *Left habenula responses.* Extracted BOLD responses from the left habenula corresponding to the dynamically changing value of win, loss and shock CSs, averaged across subjects. The linear response profile is similar to the right habenula, but not statistically significant.

Fig. S3: *Medial dorsal thalamus results*. No significant BOLD response to the value of win, loss and shock CSs was detected in the combined left and right MD thalamus ROI. Error bars represent SEM.

Fig. S4: *Habenula response to tonic cue value.* Extracted habenula BOLD response (averaged across left and right) corresponding to the contrast of high *versus* low probability (H-L) win, lose and shock CSs. Error bars represent SEM.

Fig. S5: *Habenula ROI results at different learning rates.* Similar results were obtained when using learning rates of α=0.3 and α=0.7 (original α=0.5). Error bars represent SEM.

Fig. S6: increased coupling between the right habenula and BA25 and posterior orbitofrontal cortex (left and middle panels) and the amygdala (right panel) as a function of increasing shock CS value. These regions were not included in our *a priori* ROIs and are presented for information only. Image thresholded at exploratory level, P < 0.005, cluster threshold (k) > 10] and overlaid on the average normalized structural image; color bars represent t-values.