Separate mesocortical and mesolimbic pathways encode effort and reward learning signals

Tobias U. Hauser, Eran Eldar & Raymond J. Dolan

Supplementary Information
**Supplementary Task Details**

We calculated the median effort employed during the red frame phase as the effort exerted at that trial. The individual effort trajectories were monitored online to ensure that the subjects kept the force approximately constant during the whole effort execution phase (feedback was given after each block if necessary). To maintain the subjects’ motivation and engagement, the points gained were converted and added to a blue bar shown at the bottom of the screen. Every time the bar reached the yellow target line, subjects received £1.00, and the bar started over from the left side of the screen, similar as implemented in previous studies (e.g., 1). Each point that the subjects won translated into an approximate 2% increase in the bar. The subjects earned £4.59±0.82 on average. To de-correlate outcome success (effort above/below threshold) from actual amount won, subjects received 0 points in 50% of all trials (probabilistically determined, cf. Fig. 1B). A jittered fixation cross (mean 6000ms, range 2000-10000ms, uniform distribution) was shown between two trials. Each of the two stimuli was presented 80 times, equally distributed across the 4 sessions of approximately 15 minutes each. Timing of task events were determined by maximizing the design efficacy for effort and reward PEs.

To measure effort, we used a bespoke, MR-compatible, pneumatic force gripper for the right hand. Air pressure was converted to digital signals using a National Instruments data converter (NI-6009, National Instruments Corporation Ltd., Newbury, UK) using a sampling rate of 1000Hz. During effort execution, the position of the thermometer was updated every 20ms. The effort feedback, as well as the threshold, was shown relative to the maximal force that the subject was able to execute. The calibration of maximum force was done at the beginning of the experiment. To ensure subjects were not deliberately squeezing below their maximum force, we determined the maximum executed force during gripper practice and replaced the maximal force if necessary. To account for slow drifts in baseline pressure due of a change in temperature that affects air expansion in the pipes, we adjusted the baseline pressure at the beginning of every scanning session. Pilot studies showed this is sufficient to account for these slow temperature drifts. It is well known that many modalities are perceived in a logarithmic rather than in a linear fashion (2). We thus used log-converted force measures for display and the determination of the force thresholds. Pilot investigations confirmed that log-transformed force feedback felt more natural and was easier to handle than non-converted feedback. Pilot studies revealed that an effort execution phase of 5000ms was feasible for males who, unlike females, were able maintain force levels across the whole experiment. Because pilot studies showed a high variability among female subjects, we decided to only test males in this study. We decided to modulate effort threshold based on the force and to keep time of execution constant, because effort is otherwise confounded by the time that one spends on the task, and thus confounds effort execution with temporal discounting (3).

Before entering the MRI, subjects were trained on the task. In a first phase, subjects were familiarized with the force grippers and learned how to control the ‘thermometer’. This task was repeated in the scanner during the survey scans so that the subjects could adjust to the new environment. During the task instructions, subjects were told about the changing force thresholds and reward magnitudes. During two practice runs, subjects learned how thresholds and points change and how they can adjust their force accordingly.

The task was delivered using the Cogent toolbox for Matlab (R2010, MathWorks, Natick MA, USA). Pulse, breathing and eye-tracking was recorded to monitor the subjects’ states and was used for artefact correction.

**Computational Modelling**

We developed a novel computational reinforcement learning model (4) predicting executed effort at each trial to understand the processes underlying the effort and reward learning in this task. The model consists of three distinct parts (Fig. S1): effort learning module, reward learning module, and a reward-effort discounting and subjective utility module.
Figure S1. Computational framework for effort and reward learning. (A) Expected rewards $\hat{R}_{t+1}$ are learned using a reward PE $\delta^R_t$ and a learning rate $\gamma$. (B) Learning about effort resembles an iterative logistic regression of beliefs about being successful given the exerted effort $p(O_t | E_t)$ (blue lines depict different beliefs about the effort threshold). This belief is then updated on each trial using an effort PE $\delta^E_t$ that adjusts the indifference point of the belief $\omega_t$. The free parameters $\alpha$ and $k$ depict a learning rate and the precision of beliefs, respectively. (C) Effort discounting of reward magnitude follows a sigmoidal function so that rewards are more strongly discounted with higher effort. Model comparison reveals that reward magnitude not only affects the height of the sigmoid but also its indifference point, which means that lower reward are discounted already at lower effort levels, whereas high rewards are discounted only with the highest effort (red lines depict how different reward magnitudes are discounted as a function of effort). (D) Based on effort and reward expectation as well as reward discounting, we can compute the subjective net benefit for each trial. The upper panel depicts how subjective net benefit changes as a function of expected reward, given a fixed belief that the effort requires is at 60%. This shows how for 1 point, the maximal benefit is just above 60%, whereas for high rewards, it is well above 60% to ensure that the points are won, accounting for the uncertainty of the effort belief. The lower panel shows how subjective net benefit changes for a reward magnitude of 3 as a function of different beliefs about the height of the effort threshold (blue lines). An example trajectory of maximal subjective net benefit can be found in Fig. 1B (pink line).

Effort learning

In the task subjects are not explicitly told how much effort is needed to succeed in a given trial, but they had to learn it based on their prior experience with a given stimulus. Because the subjects do not receive explicit feedback about the height of the effort threshold, it is not possible to calculate the effort prediction error in a simple Rescorla-Wagner-like (5) fashion, i.e. as the difference between expected and received effort (cf. eq. (1.5)). Rather, the subjects will update their belief about the probability of succeeding $p(O_t | E_t)$ on every trial $t$. To do so, we used a modified version of an iteratively reweighted least squares logistic regression (IRLS, Fig. S1B; 6). Here, the belief of succeeding is computed using a sigmoidal transformation of the executed effort at trial $t$ $E_t$

$$p(O_t | E_t; \omega_t, k) = \frac{1}{1 + e^{-k(E_t - \omega_t)}}, \quad (1.1)$$

where $\omega_t$ describes the indifference point of the sigmoid, which is equivalent to the current belief about the height of the effort threshold. Parameter $k$ describes the uncertainty about the height and is used as a free parameter. $E_t$ is the median log-force that was exerted between 0 (no force exerted) and 100 (individual maximum force, determined during practice). To account for additional perceptual uncertainty about the effort actually exerted (because the median over 4 seconds was used, we do not assume perfect knowledge), we converted $E_t$ into a Gaussian distribution with a standard deviation of 5 (arbitrarily chosen) around the mean of the actually exerted force using sampling from a normal distribution (100 samples). For outcome, the actual outcome was used assuming no uncertainty about the visual feedback).
On every trial, \( \omega_t \) is updated using an effort prediction error \( \delta^E_t \) and a learning rate \( \alpha \) (free parameter):

\[
\omega_{t+1} = \omega_t - \alpha \delta^E_t
\]

(1.2)

The effort prediction error \( \delta^E_t \) is the difference between the success at trial \( t \), \( O_t \), and the prior belief given the exerted force \( p(O_t \mid E_t; \omega_t, k) \):

\[
\delta^E_t = O_t - p(O_t \mid E_t; \omega_t, k)
\]

(1.3)

This learning rule can be seen as a simplified version on a temporal-difference based predictive learner (7), only that we did not incorporate the gradient of our prediction (cf eq. 2 in (7)) because our trajectories were relatively smooth and thus a gradient would have little effect (i.e. act as a scaling factor, which is now absorbed in learning rate \( \alpha \)). Moreover, it is difficult to imagine how a gradient was effectively implemented in a biological system such as the human brain.

**Reward learning**

To learn about the number of points that are associated with a stimulus, we used a standard Rescorla-Wagner learning model (5) (Fig. S1A), where the expected reward magnitude \( \hat{R} \) is being updated using a reward prediction error \( \delta^R_t \) and a learning rate \( \gamma \) (free parameter):

\[
\hat{R}_{t+1} = \hat{R}_t + \gamma \delta^R_t
\]

(1.4)

The reward prediction error is the difference between the expected \( \hat{R}_t \) and number of points that was presented \( R_t \):

\[
\delta^R_t = R_t - \hat{R}_t
\]

(1.5)

In this study, we used the current number shown on the screen as \( R_t \), irrespective of whether the subject exerted enough effort to surpass the effort threshold. We decided to do so, because subjects learn about the reward magnitude irrespective of their effort success, and thus update their reward expectations. Moreover, model comparison (see main text) revealed that subjects also learned about probabilistic (50%) 0-outcomes, rather than ignoring the magnitude at that trial.

**Reward-effort discounting and subjective net benefit**

It is well known that effort and other costs discount rewards (8) and recent studies investigated the effort-discounting function in great detail (9–11). Here, we compared previously suggested functions (hyperbolic, quadratic, sigmoidal) and extended these. For all models, we introduced a utility parameter \( \tau \) that accounts for non-linearities in the subjective representation of reward (9, 12–14) and this improves model fit (Fig. S2).

The quadratic discounting calculated the subjective value of a reward \( v(R|e) \) given an effort \( e \), is based on Hartman et al.’s studies (10, 11), and has a free decay kernel parameter \( \kappa \) that depicts the discounting steepness:

\[
v(R|e) = R^\tau - \kappa e^\tau
\]

(1.6)

The hyperbolic discounting was originally introduced as mirroring hyperbolic temporal discounting and is formalized with a discounting kernel \( \kappa \):

\[
v(R|e) = \frac{R^\tau}{1 + \kappa e^\tau}
\]

(1.7)

The sigmoidal effort-discounting function is more flexible than the quadratic and hyperbolic functions and has two free parameters: indifference point \( \rho \) and the slope \( \kappa \). Similar to Klein-Függe et al. (9), we extended the sigmoid by two terms: subtracting \( \frac{1}{1 + e^{\kappa p}} \) to ensure that \( v(R) = R^\tau \) when effort \( e \) equals 0 (i.e., no
discounting). By multiplying \( 1 + \frac{1}{e^{\kappa p}} \), it is ensured that the subjective values \( v(R) \) will not become negative for high effort, in keeping with previous modelling of effort discounting (9).

\[
v(R | e) = R^T \left( 1 - \frac{1}{1 + e^{-\kappa (e - p)}} - \frac{1}{1 + e^{\kappa p e}} \right) \left( 1 + \frac{1}{e^{\kappa p}} \right)
\]

(1.8)

In this sigmoidal discounting, the reward only affects the height of the sigmoid, but the indifference point \( p \) is unaffected by the reward. This means that the acceleration of discounting always occurs at the same effort level. However, it might be that with high rewards, discounting only emerges at very high effort, whereas low rewards discount at low effort (cf. horizontal shifts in Fig. S1). We thus implemented an additional sigmoidal function where height and indifference point are modulated by expected rewards:

\[
v(R | e) = R^T \left( 1 - \frac{1}{1 + e^{-\kappa (e - p R')}} - \frac{1}{1 + e^{\kappa p R'}} \right) \left( 1 + \frac{1}{e^{\kappa p R'}} \right)
\]

(1.9)

On each trial, we assume that subjects exert the effort that has the highest subjective net benefit for them by combining the beliefs about the effort threshold and the belief about the current reward. The net benefit for every effort \( e \) (0-100%, Fig. S1) is calculated as the product of the belief about the effort threshold and the subjective value of that effort, and the normalized probability density at the chosen effort was the used for calculating the model fit (log likelihood):

\[
B(e) = p(O_t | \epsilon; \omega_t, k) v(\hat{R}_t | e)
\]

(1.10)

Alternatively, one could additionally account for the utility of not succeeding (i.e. expending effort without receiving a reward). Because the probability of not succeeding in this task is the inverse of the success-probability, such a model results in almost identical results as simulations revealed. We thus decided to compute utility as in eq. (1.10), similar to previous studies (9, 11).

Model parameters were estimated by maximizing the probabilistic benefit function, independently for each subject, using a genetic algorithm (15), and model comparison was done using fixed- and random-effects analysis for BIC (16).

Please note that the goal of our modelling was to derive a model that most adequately reflects the learning signals, and we did not seek to design it in order to maximize orthogonality between free parameters. It is thus possible that some of the parameters express a considerable covariance, and we thus decided to not analyze the parameters in detail.

Model comparison and selection

We compared several potential models and the best fitting model was selected for further analysis (Fig. S2). First, we compared the model fit using the four different effort-discounting functions. The sigmoidal effort-discounting, where height and indifference point are modulated by reward outperformed the other models clearly (eq. (1.9)).

Further model comparisons revealed that by imposing no learning for either rewards or effort (by setting the expected reward/effort to the mean expectation across the task), the model fits were clearly worse. This confirms that subjects learned about rewards and effort, and that these learning processes are necessary for the model to explain the behavior.

In addition, we compared the effort PE model to a heuristic effort learner. This model adjusted its effort expectations based on the outcome (success/no success). However, in contrast to the effort PE model, it did so by merely adjusting its expectation by the same amount, stable across trials. This means that such a model ignores the size of a prediction error (eq. 1.2) and only uses its valence, and consequently always changes the effort expectation by the same amount.

We also compared our model to a model in which rewards are learned using (near) optimal inference. In this task, the best approximation of the current reward magnitude is achieved by taking the previously displayed
reward magnitude while ignoring the probabilistic 0 rewards. This optimal model performed better than the no-learning model, but worse than the PE-based learning model.

Moreover, a model ignoring stimulus identity performed worse than the winning model, thus supporting the notion that subjects take stimulus-identity into account (logL=-17320, BIC=35523).

Lastly, we found that a model without utility parameter τ performs worse and a reward learning algorithm that ignores the probabilistic 0 reward magnitudes also has slightly worse model fits.

**Figure S2.** Model comparison. Model comparison revealed that a sigmoidal effort-discounting function, where height and indifference point are modulated by reward (‘sigmoidal + height-modulation’) performs best. Models that do not employ PE-based learning, such as a no-learning model, use heuristics or optimal inference model perform worse. Both, a utility parameter τ and the learning of 0-reward trials turned out to improve model fits. The model winning model comparison (using BIC) in a fixed-effects (middle panel) and a random effects analysis (right panel; , 17) is highlighted in bright gray.
Figure S3. Salience PEs across domains. Whole-brain conjunction analysis (p<.05, cluster-extent FDR correction, height threshold p<.001) reveals that unsigned, salience PEs of both domains activate a common network including the left anterior insula (A; MNI: -29, 20 -6, cluster 106 voxels, peak t=4.48) and intraparietal sulcus (B; MNI: -47 -62 44, cluster 179 voxel, peak t=4.35).

Figure S4. Correlation between fMRI regressors. To control for potential collinearities between our regressors of interest, we disabled the orthogonalization in our analysis, letting the regressors compete for variance (18). Because we independently varied effort and reward trajectories, we were able to achieve very low correlations between our regressors. Color bar: average Pearson correlation coefficient; rPE: reward prediction error (PE); ePE: effort PE; abs rPE: absolute, salience rPE; abs ePE: salience ePE.
Figure S5. Double dissociation of reward and effort predictions error in cortex and striatum. The double-dissociation between effort and reward PEs was also evident in literature-based regions-of-interest of (A) the ventral striatum (VS-ROI derived from http://www.neurosynth.org/; reward PE: t(27)=6.93, p<.001; effort PE: t(27)=.25, p=.807; comparison: t(27)=3.74, p<.001), and (B) the dmPFC (ROI derived from (19); MNI: 2 18 54, 10mm sphere; effort PE: t(27)=5.92, p<.001; reward PE: t(27)=1.31, p=.202; comparison: t(27)=-2.70, p=.012). Light blue indicates ROIs overlaid over main effects as shown in Fig. 2. Whole-brain comparison between reward and effort prediction errors confirmed a double-dissociation in dmPFC (C) and VS (D) on a whole-brain corrected level. (E) Our findings do not contradict previous findings showing reward PEs in the medial wall (18, 20–22), as we also found reward PEs, but in distinct, more ventral areas of the medial wall (RPE: reward prediction error).
Figure S6. Coronal view on reward and effort PEs in SN/VTA. The spatial gradients for effort and reward PE distributions confirms that reward PEs (red) are primarily processed in dorsomedial regions, whereas effort PE (blue) are represented in ventrolateral areas. Also see Figures 3 & 4.

Figure S7. Effort representations along the medial wall. Effort PE activation (warm colors, as in Fig. 2D) spans dmPFC between pre-SMA and ACC (anatomical regions in pink derived from Iannaccone et al., 2015, (23)). The effort PE activation lies in close proximity with a previous finding (19) of effort outcomes (blue). Interestingly, the effort PEs lie anterior to activations for effort expectation (green). Both, a previous (24) and our study (effort expectation during anticipation: p<.05 cluster-extend FWE, MNI: [-15 -12 62], t=4.88, cluster size=494) find effort expectation signals in SMA. This suggests that effort evaluation is processed more anteriorly to effort expectation. All activations are projected to the same sagittal plane (x-axis) for visualization purposes.
**Table S1.** Effort and reward prediction error activation. Main effect of effort and reward prediction error. Areas shown that are significant at $p<.05$, height-FWE corrected for multiple comparisons, $k>30$, unless stated otherwise. ACC: anterior cingulate cortex; dmPFC: dorsomedial prefrontal cortex; IPS: intraparietal sulcus; ITL: inferior temporal lobe; n.s.: no significance; pre-SMA: pre-supplementary motor area; SN/VTA: substantia nigra/ventral tegmental area; VS: ventral striatum

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Region</th>
<th>Hemisphere</th>
<th>Cluster size (voxels)</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>T Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>effort prediction error</td>
<td>temporal pole</td>
<td>right</td>
<td>39</td>
<td>48</td>
<td>8</td>
<td>-18</td>
<td>8.39</td>
</tr>
<tr>
<td></td>
<td>dmPFC</td>
<td>bilateral</td>
<td>47</td>
<td>-2</td>
<td>15</td>
<td>62</td>
<td>7.45</td>
</tr>
<tr>
<td></td>
<td>(pre-SMA/ACC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SN/VTA</td>
<td>bilateral</td>
<td>6</td>
<td>-3</td>
<td>-14</td>
<td>-15</td>
<td>4.34*</td>
</tr>
<tr>
<td>-effort prediction error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reward prediction error</td>
<td>visual cortex</td>
<td>left</td>
<td>228</td>
<td>-38</td>
<td>-75</td>
<td>-8</td>
<td>11.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>right</td>
<td>818</td>
<td>-29</td>
<td>-84</td>
<td>17</td>
<td>10.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>61</td>
<td>-39</td>
<td>-72</td>
<td>14</td>
<td>7.91</td>
</tr>
<tr>
<td></td>
<td>IPS</td>
<td>right</td>
<td>1764</td>
<td>30</td>
<td>-48</td>
<td>48</td>
<td>11.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>left</td>
<td>32</td>
<td>51</td>
<td>-33</td>
<td>53</td>
<td>7.58</td>
</tr>
<tr>
<td></td>
<td>VS</td>
<td>right</td>
<td>319</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>10.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>left</td>
<td>81</td>
<td>-14</td>
<td>8</td>
<td>-3</td>
<td>8.29</td>
</tr>
<tr>
<td></td>
<td>ITL</td>
<td>right</td>
<td>336</td>
<td>41</td>
<td>-59</td>
<td>-11</td>
<td>9.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>left</td>
<td>82</td>
<td>-38</td>
<td>-57</td>
<td>-5</td>
<td>9.49</td>
</tr>
<tr>
<td></td>
<td>SN/VTA</td>
<td>left</td>
<td>109</td>
<td>-9</td>
<td>-26</td>
<td>-12</td>
<td>5.63**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>right</td>
<td>104</td>
<td>12</td>
<td>-26</td>
<td>-11</td>
<td>5.07***</td>
</tr>
</tbody>
</table>

* $p=.011$, small-volume FWE-corrected for anatomical SN/VTA mask  
** $p=.001$, small-volume FWE-corrected for anatomical SN/VTA mask  
*** $p=.002$, small-volume FWE-corrected for anatomical SN/VTA mask
References


