Opponent intracerebral signals for reward and punishment prediction errors in humans

Maëlle CM Gueguen[1], Pablo Billeke[2], Jean-Philippe Lachaux[3], Sylvain Rheims[4], Philippe Kahane[5], Lorella Minotti[5], Olivier David[1], Mathias Pessiglione[6,7,8] and Julien Bastin[1, 8]

[1] Univ. Grenoble Alpes, Inserm, U1216, Grenoble Institut Neurosciences, GIN, 38000 Grenoble, France

[2] División de Neurociencia, Centro de Investigación en Complejidad Social (neuroCICS), Facultad de Gobierno, Universidad del Desarrollo, Santiago, Chile

[3] Lyon Neuroscience Research Center, Brain Dynamics and Cognition team, DYCOG INSERM UMRS 1028, CNRS UMR 5292, Université de Lyon, F-69000, Lyon, France

[4] Department of Functional Neurology and Epileptology, Hospices Civils de Lyon and University of Lyon, Lyon, France

[5] Univ. Grenoble Alpes, Inserm, U1216, CHU Grenoble Alpes, Grenoble Institut Neurosciences, GIN, 38000 Grenoble, France

[6] Motivation, Brain and Behavior lab, Centre de NeuroImagerie de Recherche, Institut du Cerveau et de la Moelle épinière, Hôpital de la Pitié-Salpêtrière, Paris, France

[7] Inserm U1127, CNRS U7225, Université Pierre et Marie Curie (UPMC-Paris 6), Paris, France

[8] These authors contributed equally

*Correspondence: julien.bastin@univ-grenoble-alpes.fr

Summary

Whether maximizing rewards and minimizing punishments rely on distinct brain systems remains debated, inconsistent results coming from human neuroimaging and animal electrophysiology studies. Bridging the gap across species and techniques, we recorded intracerebral activity from twenty patients with epilepsy while they performed an instrumental learning task. We found that both reward and punishment prediction errors (PE), estimated from computational modeling of choice behavior, correlated positively with broadband gamma activity (BGA) in several brain regions. In all cases, BGA increased with both outcome (reward or punishment versus nothing) and surprise (how unexpected the outcome is). However, some regions (such as the ventromedial prefrontal and lateral orbitofrontal cortex) were more sensitive to reward PE, whereas others (such as the anterior insula and dorsolateral prefrontal cortex) were more sensitive to punishment PE. Thus, opponent systems in the human brain might mediate the repetition of rewarded choices and the avoidance of punished choices.

Keywords

reinforcement learning; appetitive learning, aversive learning; monetary gains; monetary losses; anterior insula; prefrontal cortex; broadband gamma; Ecog; iEEG

INTRODUCTION

Approaching reward and avoiding punishment are the two fundamental drives of animal behavior. As the philosopher John Locke would put it “reward and punishment are the only motives to a rational creature: these are the spur and reins whereby all mankind are set on work, and guided”. In principle, both reward-seeking and punishment-avoidance could be learned through the same algorithmic steps. One the most straight and simple algorithm postulates that the value of chosen action is updated in proportion to prediction error (Rescorla and Wagner, 1972; Sutton and Barto, 1998), defined as observed minus expected outcome value. In this simple reinforcement learning model, the only difference is outcome valence: positive for reward (increasing action value) and negative for punishment (decreasing action value). The same brain machinery could therefore implement both reward and punishment learning.

Yet, different lines of evidence point to an anatomic divide between reward and punishment learning systems, in relation with opponent approach and avoidance motor behaviors (Boureau and Dayan, 2011; Pessiglione and Delgado, 2015). First, fMRI studies have located prediction error (PE) signals in different brain regions, such as the ventral striatum and ventromedial prefrontal cortex (vmPFC) for reward versus the amygdala, anterior insula (aINS) or lateral orbitofrontal cortex (lOFC) for punishment (O’Doherty et al., 2001; Pessiglione et al., 2006; Seymour et al., 2005; Yacubian, 2006). Second, reward and punishment learning can be selectively affected, for instance by dopaminergic manipulation and anterior insular lesion (Bodi et al., 2009; Frank, 2004; Palminteri et al., 2012; Rutledge et al., 2009).

However, a number of empirical studies have casted doubt on this anatomical separation between reward and punishment learning systems. Part of the confusion might come from the use of behavioral tasks that allow for a change of reference point, such that not winning becomes punishing and not losing becomes rewarding (Kim et al., 2006). The issue is aggravated with decoding approaches that preclude access to the sign of PE signals, i.e. whether they increase or decrease with reward versus punishment (Vickery et al., 2011). Another reason for inconsistent findings might be related to the recording technique: fMRI instead of electrophysiology. Indeed, some electrophysiological studies in monkeys have recorded reward and punishment PE signals in adjacent brain regions (Matsumoto and Hikosaka, 2009; Monosov and Hikosaka, 2012). In addition, single-unit recordings in monkeys have identified PE signals in different brain regions: not only small deep brain

nuclei such as the ventral tegmental area (Bayer and Glimcher, 2005; Schultz et al., 1997) but also large cortical territories such as the dorsolateral prefrontal cortex (dlPFC, Asaad et al., 2017; Oemisch et al., 2019).

A key issue with fMRI is that the temporal resolution might not be sufficient to dissociate the two components of PE - observed and expected outcome value. The issue arises because the same region might reflect PE at both the times of option and outcome display. Thus, if option and outcome display are close in time, the hemodynamic signals reflecting positive and negative expected outcome value would cancel each other (Roy et al., 2014; Rutledge et al., 2010). Moreover, discrepant results between human and monkey studies could also arise from other differences in the paradigms (Wallis, 2012), such as the amount of training or the particular reward and punishment items used to condition choice behavior in monkeys. Indeed, primary reinforcers such as fruit juices and air puffs may not be exact reward and punishment equivalents, as are the monetary gains and losses used in humans.

With the aim of bridging across species and techniques, we investigate here PE signals in the human brain, using a time-resolved recording technique: intracerebral electroencephalography (iEEG). The iEEG signals were collected in patients implanted with electrodes meant to localize epileptic foci, while they performed an instrumental learning task. The same approach was used in one previous study that failed to identify any anatomical specificity in the neural responses to positive and negative outcomes (Ramayya et al., 2015). To assess whether this lack of specificity was related to the recording technique or to the behavioral task, we used a task that properly dissociates between reward and punishment learning, as shown by previous fMRI, pharmacological and lesion studies (Palminteri et al., 2012; Pessiglione et al., 2006).

In this task (Figure 1a), patients (n=20) were required to choose between two cues to maximize monetary gains (during reward-learning) or minimize monetary losses (during punishment-learning). Reward and punishment PE could then be inferred from the history of tasks events, using a computational model. We first identified from the 1694 cortical recording sites a set of brain regions encoding PE, which included vmPFC, lOFC, aINS and dlPFC. We then specified the dynamics of PE signals in both time and frequency domains, and compared between reward and punishment outcomes. Results suggest a dissociation between two functionally opponent systems encoding both components (outcome minus expectation) of either reward (vmPFC and lOFC) or punishment (aINS and dlPFC) PE signals.

RESULTS

iEEG data were collected from twenty patients with drug-resistant epilepsy (see demographical details in Table S1 and methods) while they performed an instrumental learning task during which reward and punishment conditions were matched in difficulty, as the same probabilistic contingencies were to be learned.

Behavioral performance

Patients were able to learn the correct response over the 24 trials of the learning session: they tended to choose the most rewarding cue in the gain condition and avoid the most punishing cue in the loss condition (Figure 1b). Average percentage of correct choices (Figure 1c) in the reward and punishment conditions was significantly different from chance level (reward: 71.4 ± 3.2%, t_{19}=6.69, $p<3\times10^{-6}$; punishment: 71.5 ± 2.1%, t_{19}=10.02, $p<6\times10^{-9}$; difference: t_{19}=-0.03; p=0.98). Reaction times were significantly shorter in the reward than in the punishment condition (Figure 1f; reward: 700 ± 60 ms; punishment: vs 1092 ± 95 ms; difference: t_{19}=-7.02, $p<2\times10^{-6}$). Thus, patients learnt similarly from rewards and punishments, but took longer to choose between cues for punishment avoidance. This pattern of results replicate behavioral data previously obtained from healthy subjects (Palminteri et al., 2015; Pessiglione et al., 2006).

Computational modeling

To generate trial-wise expected values and prediction errors, we fitted a Q-learning model (QL) to behavioral data. The QL model generates choice likelihood via a softmax function of cue values, which are updated at the time of outcome. Fitting the model means adjusting two parameters (learning rate and choice temperature) to maximize the likelihood of observed choices (see methods). Because this simple model left systematic errors in the residuals, we implemented another model (QLr) with a third parameter that increased the value of the cue chosen in the previous trial, thereby increasing the likelihood of repeating the same choice. We found that including a repetition bias in the softmax function better accounted for the data, as indicated by a significantly lower Bayesian information criterion for QLr model (t_{19}=4.05, p<0.001; Table 1). On average, this QLr model accounts for a more symmetrical

performance between reward and punishment learning (Figure 1d), while the standard QL model would learn better in the reward condition, because reinforcement is more frequent than in the punishment condition (as patients approach the +1€ and avoid the -1€ outcome). With the QLr model, choices in reward and punishment conditions were captured equally well, with an explained variance across patients of 87 and 83% (Figure 1e).

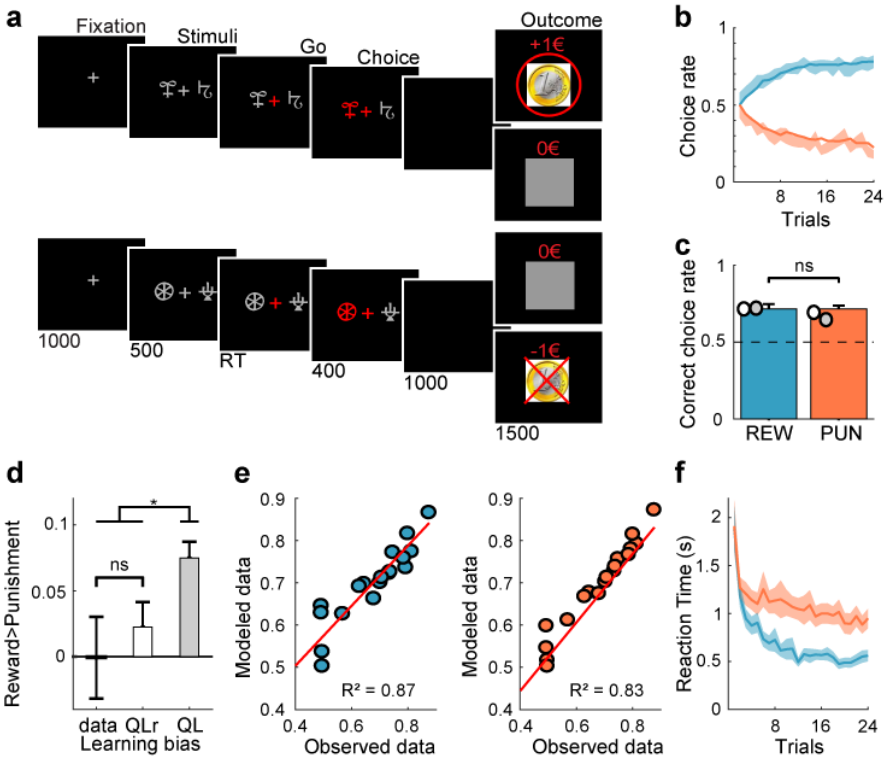

Figure 1. Behavioral task and results. **a.** Successive screenshots of a typical trial in the reward (top) and punishment (bottom) conditions. Patients had to select one abstract visual cue among the two presented on each side of a central visual fixation cross, and subsequently observed the outcome. Duration is given in milliseconds. **b.** Average learning curves (n=20 patients). Modeled behavioral choices (solid line) are superimposed on observed choices (shaded areas represent mean ± SEM across patients). Learning curves show rates of correct choice (75% chance of 1€ gain) in the reward condition (blue curves) and incorrect choice (75% chance of 1€ loss) in the punishment condition (red curves). **c.** Average performance (correct choice rate). Modeled performance is indicated by white and grey disks (using Q-learning + repetition bias and basic Q-learning model, QLr and QL, respectively). **d.** Difference between conditions (reward minus punishment correct choice rate) in observed and modeled data. **e.** Inter-patient correlations between modeled and observed correct choice rate for reward and punishment learning. Each circle represents one patient. Red line represents the linear regression. **f.** Reaction time (RT) learning curves. Median RT are averaged across patients and the mean (± SEM) is plotted as function of trials separately for the reward (blue) and punishment (red) conditions.

Table 1. Model parameters and comparison criterion.

	Degrees of freedom (DF)	Bayesian information criterion (BIC)	Learning rate (α)	Inverse temperature (β)	Repetition bias (θ)
QL	2	502 ±31	0.27 ± 0.04	3.80 ± 0.48	
QLr	3	430 ±30	0.26 ± 0.04	3.19 ± 0.43	0.44 ± 0.06

iEEG: localizing PE using broadband gamma activity

To identify brain regions signaling PE, we first focused on broadband gamma activity (BGA, in the 50-150Hz range) because it is known to correlate with both spiking and fMRI activity (Lachaux et al., 2007; Mukamel et al., 2005; Niessing, 2005; Nir et al., 2007). BGA was extracted from each recording site and time point and regressed against PE (collapsed across reward and punishment conditions) generated by the QLr model across trials. The location of all iEEG recording sites (n=1694 bipolar derivations) was labeled according to MarsAtlas parcellation (Auzias et al., 2016), and to the atlas of Destrieux (Destrieux et al., 2010) for the hippocampus and the distinction between anterior and posterior insula (Figure 2). In total, we could map 1473 recording sites into 39 brain parcels. In the following, we report statistical results related to PE signals tested across the recording sites located within a given parcel. Note that a limitation inherent to any iEEG study is that the number of recorded sites varies across parcels, which impacts the statistical power of analyses used to detect PE signals in different brain regions.

For each parcel, we first tested the significance of regression estimates (averaged over the 0.25 to 1 s time window following outcome onset) in a fixed-effect analysis (pooling sites across patients) and Bonferroni-corrected for multiple comparisons across parcels. We also estimated the significance of PE signals at the site level, by using time-varying regression estimates and associated p-values, while FDR-correcting for multiple comparisons in the time domain (across 97 comparisons in the 0 to 1.5 s time window following outcome onset), in accordance with published methods (Genovese et al., 2002). We found 8 parcels showing significant PE signals and displaying a proportion of significant contacts superior to 20% (Table S2). This set of significant brain parcels included the aINS, vmPFC, dlPFC, lOFC, hippocampus, lateral and caudal medial visual cortex (VCcm and VCl) and the medial inferior temporal cortex (ITcm). Given this result and the literature reviewed in the introduction, we focused on the anterior insula and prefrontal ROIs (vmPFC, lOFC and dlPFC) in the hereafter analyzes, while the same analyzes performed in the other ROIs are analyzed are presented as supplementary information.

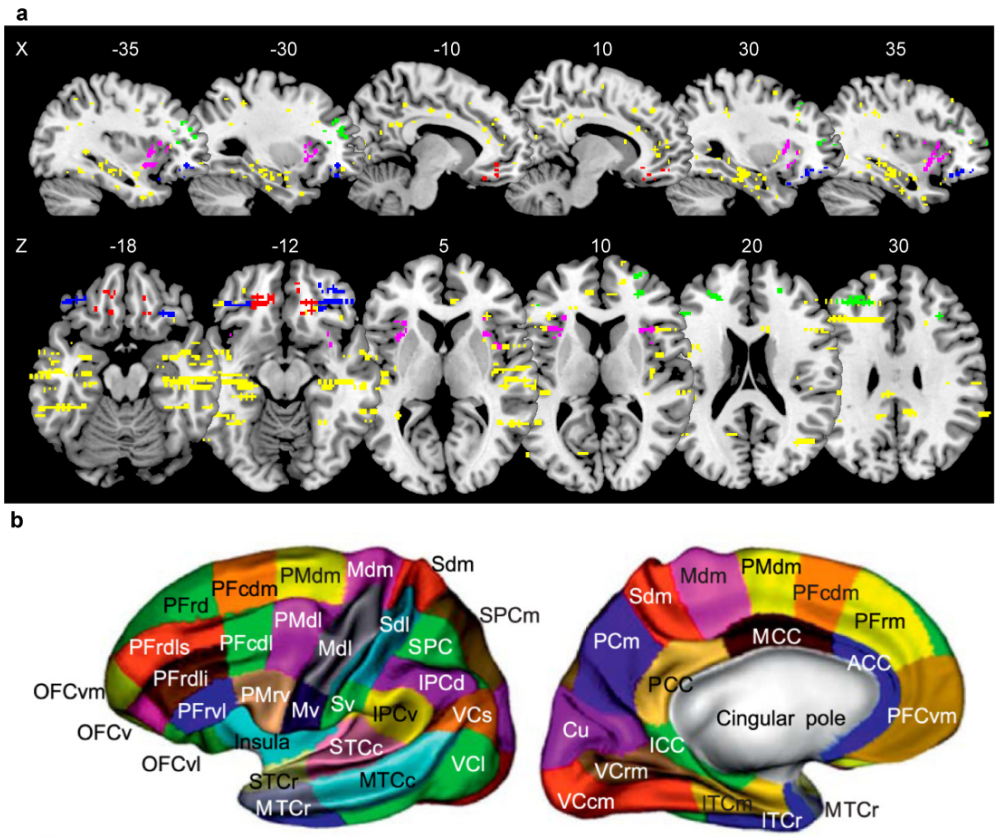

Figure 2. Anatomical location of intracerebral electrodes. a. Sagittal and axial slices of a brain template over which each dot represents one iEEG recording site (n=1694). Color code indicates location within the four main regions of interest (red: vmPFC; green: dlPFC; blue: lOFC; purple: aINS). **b.** MarsAtlas parcellation scheme represented on an inflated cortical surface.

iEEG: PE signals across ROIs and frequency bands

In each ROI (Figure 3a), we explored whether activity in other frequency bands could also be related to PE. A time-frequency analysis confirmed the presence of PE signals in BGA following outcome onset in all ROIs (Figure 3b). Furthermore, PE was also positively associated with beta-band (13-33 Hz) power in the aINS and vmPFC. In the theta/alpha bands (4-8 and 8-13 Hz), there was an initial positive association (during the first 500 ms after outcome onset), which was followed by a negative association (from 500 ms to 1000 ms after outcome onset) in all four ROIs. Thus, the time-frequency analysis pointed to three other frequency bands in which power was associated to PE.

We regressed trial-wise power against PE, in the four ROIs and four frequency bands, for each time point between -0.2 and 1.5 s around outcome onset. In the broadband gamma (Figure 3c), we confirmed a significant cluster-corrected association with PE in the 0.09-1.00s window for the aINS ($sum(t_{(82)})=462.2$, $p_c<1\times10^{-3}$), 0.19-0.97s for the dlPFC ($sum(t_{(73)})=273.5$, $p_c<1\times10^{-3}$), 0.30-0.86s for the vmPFC ($sum(t_{(53)})=115.3$, $p_c<1\times10^{-3}$) and 0.39-0.94s for the lOFC ($sum(t_{(69)})=116.1$, $p_c<1\times10^{-3}$). We next focused on a 0.25-1s post-outcome time window for subsequent analyses (Figure 3c), as it plausibly corresponds to the computation of PE.

To further quantify statistically how the information about PE was distributed across frequencies, we averaged regression estimates over the 0.25-1s time window for the broadband gamma and beta bands, and over two separate time windows to distinguish the early (0 to 0.5 s) and late (0.5 to 1 s) components of theta-alpha band activity (Figure 3d). As expected, we found significant PE correlates in BGA. Furthermore, beta-band activity was also positively associated with PE in two ROIs (aINS: $t_{(82)}=9.17$; $p<1\times10^{-13}$; vmPFC: $t_{(53)}=3.34$ $p=0.0015$). Finally, regarding the theta/alpha band, regression estimates were significantly above (below) zero in the early (late) time window in all ROIs (all p values<0.05).

To compare the contribution of activities in the different frequency bands to PE signaling across the four ROIs, we included them as separate regressors in general linear models meant to explain PE. In particular, to test whether other frequency bands were adding any information about PE, we compared GLMs including only BGA to all possible GLMs containing broadband gamma plus any combination of low-frequency activities. Bayesian model selection (see Methods) designated the broadband-gamma-only GLM as providing the best account of PE (Ef=0.997, Xp=1). Thus, even if low-frequency activity was significantly

related to PE, it carried redundant information relative to that extracted from BGA.

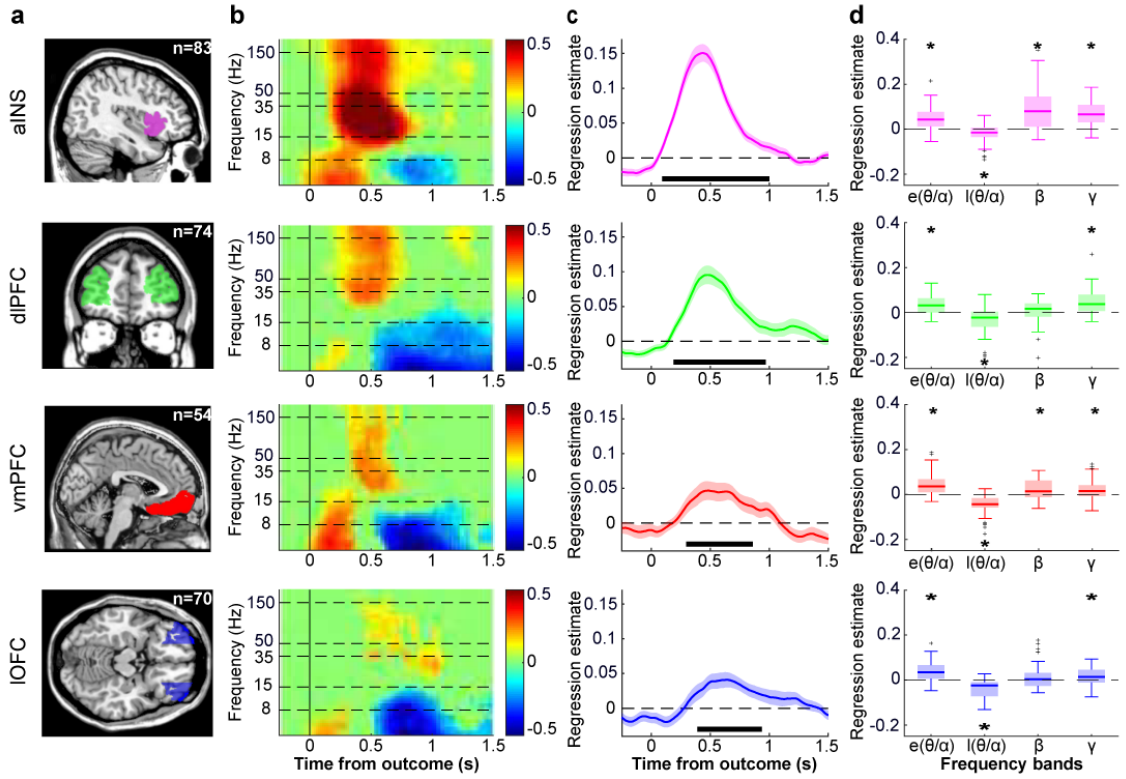

Figure 3. Investigation of PE signals across frequency bands. a. Anatomical localization of the aINS (purple), dlPFC (green), vmPFC (red) and lOFC (blue). All recording sites located in these parcels were included in the ROI analyzes. **b**. Time-frequency decomposition of PE signals following outcome onset. Hotter colors indicate more positive regression estimates. Horizontal dashed lines indicate boundaries between frequency bands that are investigated in panels c and d. **c**. Time course of regression estimates obtained from linear fit of BGA with PE modeled across reward and punishment conditions. Horizontal bold lines indicate significant clusters ($p_c<1\times10^{-3}$). **d.** Regression estimates of power against PE, averaged over early (0 to 0.5 s) and late (0.5 to 1 s) post-stimulus windows for the lower frequency bands (θ/α: 4-13 Hz) and over the 0.25 to 1 s window for higher frequency bands (β: 13-33 Hz and broadband γ: 50-150 Hz). Black stars indicate significance of regression estimates. Black crosses indicate outliers.

iEEG: comparison between reward and punishment PE

In the following analyses, we tested whether reward and punishment PE could be dissociated between the four ROIs previously identified (aINS, dlPFC, vmPFC and lOFC). The time course of regression estimates computed separately for reward and punishment PE showed increases at the time of outcome display, which differed between ROIs (Figure 4a). In aINS and dlPFC, regression estimates were significantly higher for punishment than for reward PE, in the 0.23-0.70 s window for the aINS ($sum(t_{(82)})$=-97.01, $p_c<1\times10^{-3}$) and in the 0.25-1.5 s for the dlPFC ($sum(t_{(73)})$=-368.6, $p_c<1\times10^{-3}$). An opposite pattern emerged in the vmPFC and lOFC: regression estimates were significantly higher for reward PE than for punishment PE, in the 0.48-1.02 s for the vmPFC ($sum(t_{(53)})$=116, $p_c<1\times10^{-3}$) and in the 0.72-1.45 s for the lOFC ($sum(t_{(69)})$=138.7, $p_c<1\times10^{-3}$).

We next decomposed PE signals into outcome and expected value, for both reward (R and Qr) and punishment (P and Qp) PE, to test whether the two components were related to BGA, at the time of outcome display. We observed a consistent pattern across ROIs: while the outcome component was positively correlated with BGA, the expectation component was negatively correlated with BGA (Figure 4b). To further quantify this pattern in each ROI, we tested regression estimates (averaged over a 0.25-1s post-outcome time window) of both outcome and expectation for both reward and punishment PE (Figure 4c). We found that all components of both reward and punishment PE were significantly expressed in aINS and dlPFC BGA (all p values<0.05), but only the reward PE components in the vmPFC and lOFC (all p values<0.05).

The same analyzes were performed in the four other brain regions associated with PE (Figure S1). BGA in the medial Inferior Temporal Cortex (mITC) was associated with outcomes (reward and punishments), but not with expected value, so this region would not qualify as signaling PE. The three other ROIs (Hippocampus: HPC; lateral Visual Cortex: lVC and caudal medial Visual Cortex: cmVC), showed a dissociation in time, with a short punishment PE and prolonged reward PE. When averaging the signal over the 0.25-1s post-outcome time window, there was no significant difference between reward and punishment PE in these regions. There was therefore no strong evidence for these regions to be associated with either reward or punishment learning.

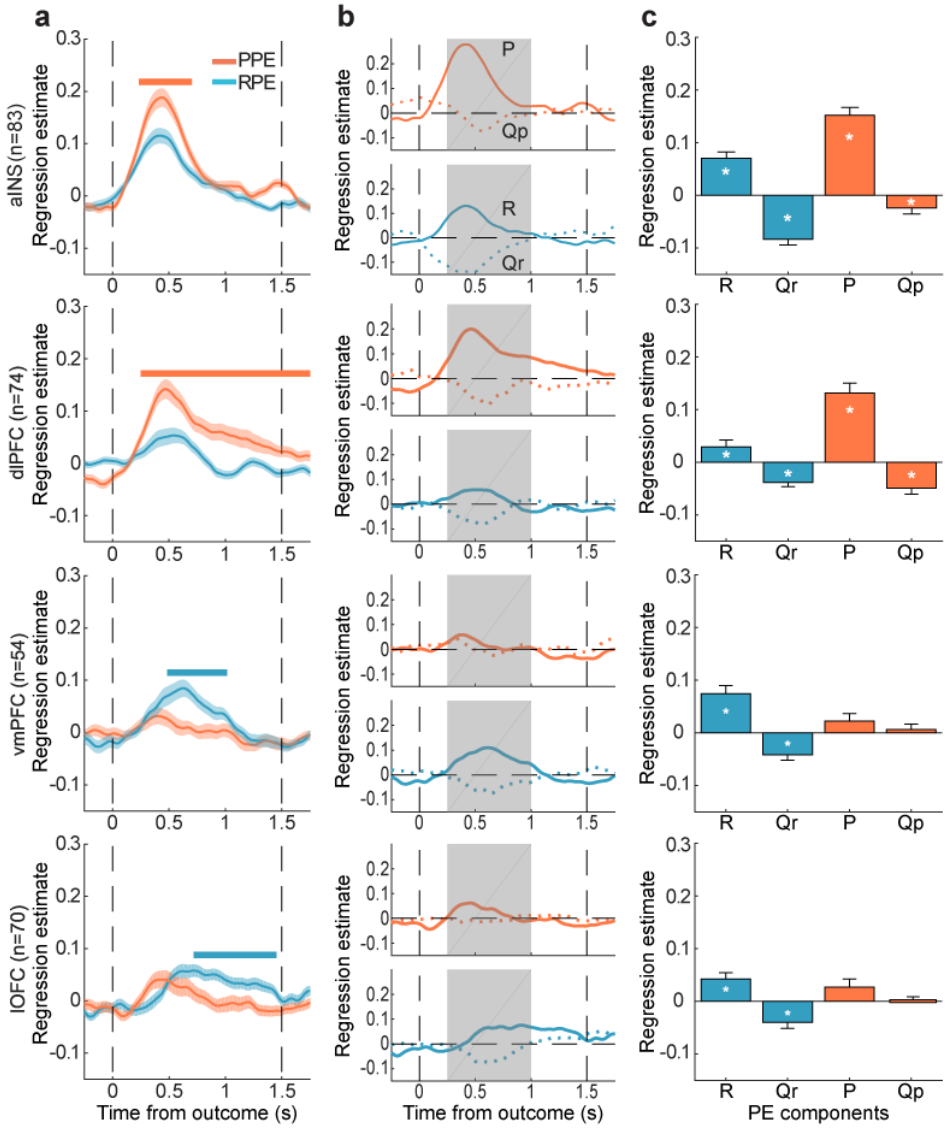

Figure 4. Dissociation of reward and punishment PE signals. a. Time course of regression estimates obtained from linear fit of BGA with PE modeled separately for the reward and punishment conditions (PPE: punishment prediction error; RPE: reward prediction error). Horizontal bold lines indicate significant difference between conditions (blue: RPE>PPE; red: PPE>RPE; $p_c<0.05$). Shaded areas represent inter-patient SEM. **b**. Time course of regression estimates obtained from a linear model including both outcome and expected value components for both reward (R and Qr) and punishment (P and Qp) PE. **c.** Regression estimates averaged over the 0.25-1 s time window (represented as shaded gray areas in panels b).

DISCUSSION

Here, we compared the neural correlates of reward and punishment PE during instrumental learning. We identified a set of brain regions signaling PE in different frequency bands, the most informative being BGA. All regions signaled outcomes with increased BGA and expectations with decreased BGA. However, there was a partial dissociation: the vmPFC and lOFC emitted stronger signals for reward PE, whereas the aINS and dlPFC emitted stronger signals for punishment PE. In the following, we successively discuss the specification of PE signals in terms of anatomical location and frequency band, and then the dissociation between reward and punishment PE.

Specification of PE signals

When regressing BGA against PE modeled across learning conditions, we identified significant correlates in a number of brain regions. Among the significant ROIs, some (e.g., the vmPFC and aINS) were classic regions related to prediction errors in meta-analyses of human fMRI studies (Bartra et al., 2013; Garrison et al., 2013; Liu et al., 2011), whereas others (e.g., the dlPFC and lOFC) were regions where single-neuron firing activity in non-human primates was shown to correlate with prediction error (Asaad et al., 2017; Oemisch et al., 2019; Sul et al., 2010). Our study thus fills a gap across species and techniques, confirming that intracerebral BGA is a relevant neurophysiological signal related to both hemodynamic and spiking activity, as previously suggested (Lachaux et al., 2012; Mukamel et al., 2005; Niessing, 2005; Nir et al., 2007).

Yet it raises the question of why fMRI studies, including those using the same task as here (Pessiglione et al., 2006), often failed to detect PE correlates in regions such as the dlPFC and lOFC. One possible explanation is the more stringent correction for multiple comparisons across voxels in fMRI studies, compared to the correction across ROIs applied here, or the absence of correction in most animal studies that typically investigate a single brain region. Another explanation would be that regions such as the dlPFC are more heterogenous across individuals, such that the group-level analyses typically conducted in fMRI studies might be less sensitive than the subject-level analysis performed here or in animal studies. Conversely, some key regions consistently found to signal PE in fMRI studies (e.g., the ventral striatum), are absent in our results, for the simple reason that they were not sampled by the electrodes implanted for clinical purposes. Even if our analysis provides some insight about the location of PE signals, it cannot be taken as a fair whole-brain analysis, since

some regions were more frequently sampled than others, biasing the statistical power and hence the sensitivity of PE detection.

In all the investigated ROIs, we also found significant links with activity in lower frequency bands. Time-frequency decomposition of PE correlates yielded remarkably similar patterns in the different ROIs, with an increase in the beta to high-gamma band, and an increase followed by a decrease in the theta to alpha band. The late decrease may reflect the opponency between theta-band activity and the other signals (broadband gamma, hemodynamic and spiking activity) that was documented in previous studies (Manning et al., 2009; Mukamel et al., 2005; Nir et al., 2007). However, the early increase is more surprising and suggests that PE are initially signaled in low-frequency activity, before reaching BGA. Yet when the different frequency bands were put in competition for predicting PE across trials, low-frequency activity proved to be redundant, with respect to the information already contained in BGA. This result is in line with our analysis of subjective valuation for decision making (Lopez-Persem et al., 2020): all the information available in neural activity could be found in BGA, even if activity in lower-frequency bands also showed significant value signals.

The timing of PE signals, peaking around 0.5 seconds after outcome onset, was roughly compatible with that observed in the hemodynamic response, which is typically delayed by 5-6 seconds. The (positive) correlation with the outcome and the (negative) correlation with the expectation were simultaneously observed. This double observation, made possible here by the high temporal resolution of iEEG, is rarely reported in fMRI studies (Fouragnan et al., 2018). One reason is that the hemodynamic response, because of its low temporal resolution, may confound positive expectation at cue onset and negative expectation at outcome onset, unless the two events are separated by a long delay (as in, e.g., Behrens et al., 2008). Our double observation corroborates a previous study showing that the differential response to positive and negative feedbacks, recorded with intracranial electrodes, is modulated by reward expectation (Ramayya et al., 2015). As the response to outcome can be viewed as an indicator of valence, and the modulation by expectation as an effect of surprise, it shows that valence and surprise can be represented in the same brain region, in accordance with the very notion of prediction error signal.

Dissociation between reward and punishment PE

Although all our ROIs exhibited a same pattern of response, their quantitative sensitivity to reward and punishment PE could be dissociated. This anatomic divide between opponent

learning systems in the brain may appear at variance with a previous fMRI study reporting that rewards and punishments are ubiquitously represented all over the brain (Vickery et al., 2011). However, this observation was made in a choice task (matching pennies) where the outcome is either reward or punishment. Thus, it is understandable that both reward and punishment regions were mobilized by the outcome in this task, since being rewarded is not being punished and vice-versa. Here, PE signals were defined by the comparison between reward or punishment outcomes and their omission, not with each other, which enabled a dissociation.

Besides, the previous conclusion was based on the finding that the information about reward versus punishment outcomes could be recovered from many brain regions. Had we applied the same decoding analysis here, we would have reached the same conclusion: information about reward versus punishment PE could be recovered in all our ROIs, precisely because their response depended on outcome valence. In other words, the contrast between reward and punishment PE was significant in all ROIs, but the critical difference between ROIs is the sign of this contrast: positive in the vmPFC and lOFC but negative in the aINS and dlPFC. The dissociation between regions signaling reward and punishment PE may also seem at odds with single-unit recordings showing reward and punishment PE signals can be found in neighboring neurons (Matsumoto and Hikosaka, 2009; Monosov and Hikosaka, 2012). Yet it should be emphasized that the dissociation observed here was only partial, compatible with the possibility that some regions contained more reward-sensitive neurons and others more punishment-sensitive neurons, even if both types can be found in all regions.

An important conclusion of our analyzes is that the dissociation was made between reward and punishment PE, not between positive and negative PE. Indeed, some learning models assume that positive and negative PE are processed differently, yielding different learning rates (e.g., Frank et al., 2007; Lefebvre et al., 2017). A strict dissociation between positive and negative PE (across valence) would imply that regions signaling reward PE with increased activity would signal punishment PE with decreased activity, and vice-versa. This would induce an ambiguity for the rest of the brain, as an omitted reward would be coded similarly to an inflicted punishment, and an avoided punishment similarly to an obtained reward. This is not the pattern that we observed: on the contrary, both reward and punishment PE were positively correlated to BGA in all regions (at least numerically, if not significantly). Yet reward and punishment PE could be distinguished by a downstream region, from the relative activity of regions more sensitive to reward and those more sensitive to punishment. Thus, rather than the sign of PE, the dissociation depended on their domain,

i.e. on whether the PE should reinforce the repetition or the avoidance of last choice.

Within the reward-sensitive regions, the vmPFC was expected, given the number of fMRI studies reporting a link between vmPFC and reward outcome, including those using the same task as here (Pessiglione et al., 2006 reanalyzed in Palminteri et al., 2012) and meta-analyses (Bartra et al., 2013; Clithero and Rangel, 2014; Garrison et al., 2013; Liu et al., 2011). The expression of reward PE in the vmPFC might relate to its position as a main efferent output of midbrain dopamine neurons, following the meso-cortical pathway (Haber and Knutson, 2010). Indeed, manipulation of dopaminergic transmission was found to interfere with reward learning, specifically (Bodi et al., 2009; Frank, 2004; Rutledge et al., 2009), an effect that was captured by reward sensitivity in a computational model of learning in this task (Pessiglione et al. 2006).The observation of reward PE signals in the lOFC was less expected, because it is generally not reported in meta-analyses of human fMRI studies and because several electrophysiology studies in animals suggested that, even if orbitofrontal cortex neurons respond to reward outcomes, they might not encode prediction errors (Roesch et al., 2010; Schultz, 2000). However, the similarity between lOFC and vmPFC reward PE signals is consistent with previous iEEG studies showing similar representation of subjective value and reward outcome in the two regions BGA (Lopez-Persem et al., 2020; Saez et al., 2018). Yet the lOFC and vmPFC reward PE signals may serve different functions, as was suggested by lesion studies in both human and non-human primates showing that the lOFC (but not the vmPFC) is critical for solving the credit assignment problem (Noonan et al., 2010, 2017).

Within the punishment-sensitive regions, the aINS was expected, as it was associated with punishment PE in our fMRI study using the same task (Pessiglione et al., 2006) and because it is systematically cited in meta-analyses of fMRI studies searching for neural correlates of punishment outcomes (Fouragnan et al., 2018; Garrison et al., 2013; Liu et al., 2011). Surprisingly, the link between aINS activity and punishment PE has seldom been explored in non-human primates. This exploration was made possible here by the development of oblique positioning techniques employed to implant electrodes, which result in a large spatial sampling of the insular cortex (Afif et al., 2010). This is important because other iEEG approaches, such as subdural recordings (Ecog), could not explore the role of the insular cortex in instrumental learning (Ramayya et al., 2015). The present result echoes a previous finding, using the same technique, that aINS BGA signals mistakes in a stop-signal task (Bastin et al., 2017). By comparison, punishment PE signals in the dlPFC were less expected, since they were not observed in fMRI results using the same task, even if it is not

uncommon to observe dlPFC activation following punishment outcomes (Fouragnan et al., 2018; Garrison et al., 2013; Liu et al., 2011).

Reward PE signals were also observed in both aINS and dlPFC regions, albeit with a lesser sensitivity. This may be interpreted as an effect of saliency rather than PE (Metereau and Dreher, 2013; Rutledge et al., 2010), as punishments were less frequent in the task than rewards (because patients learned to avoid the former and obtain the latter). However, pure saliency coding would not explain the responses to punishments observed in the aINS during Pavlovian learning tasks where high punishments were controlled to be more frequent than low punishments (e.g., Seymour et al., 2004) or in gambling tasks where punishment and reward outcomes were matched (e.g., Petrovic et al., 2008). Also, saliency coding would not predict the consequence of aINS damage, which was found to specifically impair punishment learning in this task, an effect that was captured by a specific diminution of the sensitivity to punishment outcome in a computational model (Palminteri et al., 2012). Yet it remains that reward and punishment learning are not exact symmetrical processes, since positive reward PE favors repetition of the same choice, whereas positive punishment PE pushes to the alternative choice, hence involving an additional switching process. This switching process might explain the longer choice RT observed in the punishment condition. The switch might relate to the prolonged implication of the dlPFC following punishment PE, in keeping with the established role of this region in cognitive control (Botvinick and Braver, 2015; Koechlin and Hyafil, 2007; Miller and Cohen, 2001). The implication of the aINS might be more related to the aversiveness of punishment PE, in line with the role attributed to this region in pain, interoception and negative feelings (Corradi-Dell'Acqua et al., 2016; Craig and Craig, 2009; Zaki et al., 2016).

In summary, we used human intracerebral BGA to test the a priori theoretical principle that reward and punishment PE could be processed by the same brain machinery (one being the negative of the other). On the contrary, we found that both reward and punishment PE were positively correlated to BGA in all brain regions. Yet some regions amplified reward PE signals, and others punishment PE signals. Thus, the opponency between reward and punishment brain systems is not about the sign of the correlation with PE, but about the valence domain of outcomes (better or worse than nothing). These appetitive and aversive domains correspond to different behaviors that must be learned: more or less approach for reward PE and more or less avoidance for punishment PE. Further research is needed to disentangle the roles of the different reward and punishment regions in these learning processes.

Methods

Patients

Intracerebral recordings were obtained from 20 patients (33.5 ± 12.4 years old, ten females, see demographical details in Table S1) suffering from pharmaco-resistant focal epilepsy and undergoing presurgical evaluation. They were investigated in two epilepsy departments (Grenoble and Lyon). To localize epileptic foci that could not be identified through noninvasive methods, neural activity was monitored in lateral, intermediate, and medial wall structures in these patients using stereotactically implanted multilead electrodes (stereotactic intracerebral electroencephalography, iEEG). All patients gave written informed consent and the study received approval from the ethics committee (CPP 09-CHUG-12, study 0907) and from a competent authority (ANSM no: 2009-A00239-48).

iEEG data acquisition and preprocessing

Patients underwent intracerebral recordings by means of stereotactically implanted semirigid, multilead depth electrodes (sEEG). Five to seventeen electrodes were implanted in each patient. Electrodes had a diameter of 0.8 mm and, depending on the target structure, contained 8–18 contact leads 2-mm-wide and 1.5-mm-apart (Dixi, Besançon, France). Anatomical localizations of iEEG contacts were determined on the basis of postimplant computed tomography scans or postimplant MRI scans coregistered with preimplant scans (Deman et al., 2018). Electrode implantation was performed according to routine clinical procedures, and all target structures for the presurgical evaluation were selected strictly according to clinical considerations with no reference to the current study.

Neuronal recordings were conducted using an audio–video-EEG monitoring system (Micromed, Treviso, Italy), which allowed simultaneous recording of 128 to 256 depth-EEG channels sampled at 256 Hz (1 patient), 512 Hz (6 patients) or 1024 Hz (12 patients) [0.1–200 Hz bandwidth]. One of the contacts located in the white matter was used as a reference. Each electrode trace was subsequently re-referenced with respect to its direct neighbor (bipolar derivations with a spatial resolution of 3.5 mm) to achieve high local specificity by cancelling out effects of distant sources that spread equally to both adjacent sites through volume conduction (Lachaux et al., 2003).

Behavioral task

Patients performed a probabilistic instrumental learning task adapted from previous studies (Palminteri et al., 2012; Pessiglione et al., 2006). Patients were provided with written instructions, which were reformulated orally if necessary, stating that their aim in the task was to maximize their financial payoff and that to do so, they had to consider reward-seeking and punishment-avoidance as equally important (Figure 1). Patients performed short training sessions to familiarize with the timing of events and with response buttons. Training procedure comprised a very short session, with only two pairs of cues presented on 16 trials, followed by 2 to 3 short sessions of five minutes, such that all patients reached a threshold of 70 % correct choices during both the reward and punishment conditions. During iEEG recordings, patients performed three to six test sessions. Each session was an independent task containing four new pairs of cues to be learned. Cues were abstract visual stimuli taken from the Agathodaimon alphabet. Each pair of cues was presented 24 times for a total of 96 trials. The four cue pairs were divided in two conditions (2 pairs of reward and 2 pairs of punishment cues), associated with different pairs of outcomes (winning 1€ versus nothing or losing 1€ versus nothing). The reward and punishment conditions were intermingled in a learning session and the two cues of a pair were always presented together. Within each pair, the two cues were associated to the two possible outcomes with reciprocal probabilities (0.75/0.25 and 0.25/0.75). On each trial, one pair was randomly presented and the two cues were displayed on the left and right of a central fixation cross, their relative position being counterbalanced across trials. The subject was required to choose the left or right cue by using their left or right index to press the corresponding button on a joystick (Logitech Dual Action). Since the position on screen was counterbalanced, response (left versus right) and value (good versus bad cue) were orthogonal. The chosen cue was colored in red for 250 ms and then the outcome was displayed on the screen after 1000 ms. In order to win money, patients had to learn by trial and error the cue–outcome associations, so as to choose the most rewarding cue in the gain condition and the less punishing cue in the loss condition.

Behavioral analysis

Percentage of correct choice (i.e., selection of the most rewarding or the less punishing cue) and reaction time were used as dependent variables. Statistical comparisons between reward and punishment learning were assessed using two-tailed paired t-tests. All statistical analyses were performed with MATLAB Statistical Toolbox (MATLAB R2017a, The MathWorks, Inc., USA).

Computational modeling

A standard Q-learning algorithm (QL) was used to model choice behavior. For each pair of cues A and B, the model estimates the expected value of choosing A (Qa) or B (Qb), according to previous choices and outcomes. The initial expected values of all cues were set at 0, which corresponded to the average of all possible outcome values. After each trial (t), the expected value of the chosen stimuli (say A) was updated according to the rule $Qa(t+1) = Qa(t) + \alpha*\delta(t)$. The outcome prediction error, $\delta(t)$, is the difference between obtained and expected outcome values, $\delta(t) = R(t) - Qa(t)$, with R(t) the reinforcement value among -1€, 0€ and +1€. Using the expected values associated with the two possible cues, the probability (or likelihood) of each choice was estimated using the softmax rule: $Pa(t) = \exp[Qa(t)/\beta] / \{\exp[Qa(t)/\beta] + \exp[Qb(t)/\beta]\}$. The constant parameters α and β are the learning rate and choice temperature, respectively. A second Q-Learning model (QLr) was implemented to account for the tendency to repeat the choice made on the preceding trial, irrespective of the outcome. A constant (θ) was added in the softmax function to the expected value of the option chosen on the previous trial presented the same cues. For example, if a subject chose option A on trial t, $Pa(t+1) = \exp[(Qa(t)+\theta)/\beta] / \{(\exp[Qa(t)+\theta)/\beta] + \exp[Qb(t)/\beta]\}$. We optimized model parameters by minimizing the negative log likelihood (LLmax) of choice data using MATLAB fmincon function, initialized at multiple starting points of the parameter space, as previously described (Khamassi et al., 2015). Bayesian information criterion (BIC) was computed for each subject and model (BIC = log(ntrials) * (ndegrees of freedom) + 2*LLmax). Outcome prediction errors (estimated with the QLr model) for each patient and trial were then Z-scored and used as statistical regressors for iEEG data analysis

Electrophysiological analysis

Collected iEEG signals were analyzed using Fieldtrip (Oostenveld et al., 2011) and homemade MATLAB codes. Anatomical labeling of bipolar derivation between adjacent contact-pairs was performed with IntrAnat software (Deman et al., 2018). The 3D T1 pre-implantation MRI gray/white matter was segmented and spatially normalized to obtain a series of cortical parcels using MarsAtlas (Auzias et al., 2016) and the Destrieux atlas (Destrieux et al., 2010). 3D models of electrodes were then positioned on post-implantation images (MRI or CT). Each recording site (i.e., each bipolar derivation) was thus labeled according to its position in a parcellation scheme in the patients' native space.

Regions of interest definition. The vmPFC ROI (54 sites) was defined as the ventromedial

PFC plus the fronto-medial part of orbitofrontal cortex bilaterally (MarsAtlas labels: PFCvm plus mesial part of OFCv and OFCvm). The lOFC ROI (n=70 sites) was defined as the bilateral central and lateral parts of the orbitofrontal cortex (MarsAtlas labels: OFCvl plus lateral parts of OFCv). The dlPFC ROI (n=74 sites) was defined as the inferior and superior bilateral dorsal prefrontal cortex (MarsAtlas labels: PFrdli and PFrdls). The aINS ROI (n=83 sites) was defined as the bilateral anterior part of the insula (Destrieux atlas labels: Short insular gyri, anterior circular insular sulcus and anterior portion of the superior circular insular sulcus).

Computation of single-trial broadband gamma envelopes. Broadband gamma activity (BGA) was extracted with the Hilbert transform of iEEG signals using custom MATLAB scripts as follows. iEEG signals were first bandpass filtered in 10 successive 10-Hz-wide frequency bands (e.g., 10 bands, beginning with 50–60 Hz up to 140–150 Hz). For each bandpass filtered signal, we computed the envelope using standard Hilbert transform. The obtained envelope had a time resolution of 15.625 ms (64 Hz). Again, for each band, this envelope signal (i.e., time-varying amplitude) was divided by its mean across the entire recording session and multiplied by 100 for normalization purposes. Finally, the envelope signals computed for each consecutive frequency bands (e.g., 10 bands of 10 Hz intervals between 50 and 150 Hz) were averaged together, to provide one single time-series (the BGA) across the entire session, expressed as percentage of the mean. This procedure was used to counteract a bias toward the lower frequencies of the frequency interval induced by the 1/f drop-off in amplitude. Finally, these time-series were smoothed with a 250 ms sliding window to increase statistical power for inter-trial and inter-individual analyses of BGA dynamics.

Computation of envelopes in lower frequencies. The envelopes of theta, alpha and beta bands were extracted in a similar manner as the broadband gamma frequency except that steps were 1 Hz for theta and alpha and 5 Hz for beta. The ranges corresponding to the different frequency bands were as follows: broadband gamma was defined as 50-150 Hz, beta as 13-33 Hz, alpha as 8-13 Hz and theta as 4-8 Hz.

Time-frequency decomposition. Time-frequency analyses were performed with the FieldTrip toolbox for MATLAB. A multitapered time-frequency transform allowed the estimation of spectral powers (Slepian tapers; lower frequency range: 4-32Hz, 6 cycles and 3 tapers per window; higher frequency range: 32-200Hz, fixed time-windows of 240ms, 4 to 31 tapers per window). This approach uses a steady number of cycles across frequencies up to 32 Hz (time window durations therefore decrease as frequency increases) whereas for frequencies

above 32Hz, the time window duration is fixed with an increasing number of tapers to increase the precision of power estimation by increasing smoothing at higher frequencies.

General linear models. Frequency envelopes of each recording site were epoched on each trial and time locked to the outcome onset (-3000 to 1500 ms). Each time series was regressed against the variables of interest to obtain a regression estimate per time point and recording site. In all GLMs, normalized power (Y) was regressed across trials against prediction error signal PE (normalized within patients) at every time point:

$Y = \alpha + \beta \times PE,$

With β corresponding to the regression estimate on which statistical tests are conducted. PE corresponds to

- Prediction errors collapsed across reward and punishment conditions in Figure 3
- Either reward or punishment PE in Figure 4. Note that punishment PE were inverted to allow an easier comparison with reward PE.

For the percentage of recorded sites related to prediction errors within each brain parcel, significance was assessed after a correction for multiple comparison in the time domain using the false discovery rate algorithm (Genovese et al., 2002).

To assess the contribution of the different frequency bands to prediction errors, we used the following GLM:

$PE = \beta_{\lambda} \times Y(\lambda) + \beta_{\beta} \times Y(\beta) + \beta_{e\theta\alpha} \times Y(e\theta\alpha) + \beta_{l\theta\alpha} \times Y(l\theta\alpha)$

With β_{λ}, β_{β}, $\beta_{e\theta\alpha}$ and $\beta_{l\theta\alpha}$ corresponding to the regression estimates of the power time series Y in the broadband gamma, beta, early alpha-theta and late alpha-theta bands. This GLM was compared to the 8 possible alternative GLMs that combine BGA power to a single other frequency band (beta or early theta-alpha or late theta-alpha), two additional frequency bands (beta and early theta-alpha or beta and late theta-alpha or early and late theta-alpha) or all possible frequency bands (beta and early theta-alpha and late theta-alpha).

The model comparison was conducted using the VBA toolbox (Variational Bayesian Analysis toolbox; available at http://mbb-team.github.io). Log-model evidence obtained in each recording site was taken to a group-level, random-effect, Bayesian model selection (RFX-BMS) procedure (Rigoux et al., 2014). RFX-BMS provides an exceedance probability (Xp)

that measures how likely it is that a given model is more frequently implemented, relative to all the others considered in the model space, in the population from which samples are drawn.

For the separate investigation of prediction error components, two separate analyses were conducted for reward and punishment PE. For each analysis, power time-series Y was regressed against both outcome (R or P) and expectation (Qr or Qp):

$Y = \alpha + \beta_1 \times R + \beta_2 \times Q$

With β_1 and β_2 corresponding to the outcome (R or P) and expectation (Qr or Qp) regression estimates.

For all GLMs, significance of regressors was assessed using one-sample two-tailed t-test. T-values and p-values of those tests are reported in the result section. Once regions of interest were identified, significance was assessed through permutation tests within each ROI. The pairing between power and regressor values across trials was shuffled randomly 60000 times. The maximal cluster-level statistics (the sum of t-values across contiguous time points passing a significance threshold of 0.05) were extracted for each shuffle to compute a 'null' distribution of effect size across a time window of -3 to 1.5 s around outcome onset. For each significant cluster in the original (non-shuffled) data, we computed the proportion of clusters with higher statistics in the null distribution, which is reported as the 'cluster-level corrected' p_c-value

Data and code availability

The data that support the findings of this study and the custom code used to generate the figures and statistics are available from the lead contact (JB) upon request.

Acknowledgments

This work benefited from the program from University Grenoble Alpes, within the program 'Investissements d'Avenir' (ANR-17-CE37-0018; ANR-18-CE28-0016; ANR-13-TECS-0013) and from the European Union Seventh Framework Program (FP7/2007–2013) under (grant 604102, Human Brain Project) . MG received a PhD fellowship from Region Rhône-Alpes (grant ARC-15-010226801). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank all patients; the sta of the Grenoble Neurological Hospital epilepsy unit; and Patricia Boschetti,

Virginie Cantale, Marie Pierre Noto, Dominique Ho□mann, Anne Sophie Job and Chrystelle Mosca for their support.

Author Contributions

MP and JB designed the experiment. MG collected the data. JPL and OD provided preprocessing scripts and anatomical location of iEEG sites. MG, PB and JB performed the data analysis. PK and SR did the intracerebral investigation and allowed the collection of iEEG data. MG, MP and JB wrote the manuscript. All the authors discussed the results and commented on the manuscript.

Declaration of Interests

The authors declare no competing interests.

REFERENCES

Afif, A., Minotti, L., Kahane, P., and Hoffmann, D. (2010). Anatomofunctional organization of the insular cortex: A study using intracerebral electrical stimulation in epileptic patients: Functional Organization of the Insula. Epilepsia *51*, 2305–2315.

Asaad, W.F., Lauro, P.M., Perge, J.A., and Eskandar, E.N. (2017). Prefrontal Neurons Encode a Solution to the Credit-Assignment Problem. J. Neurosci. *37*, 6995–7007.

Auzias, G., Coulon, O., and Brovelli, A. (2016). *MarsAtlas*: A cortical parcellation atlas for functional mapping: MarsAtlas. Hum. Brain Mapp. *37*, 1573–1592.

Bartra, O., McGuire, J.T., and Kable, J.W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage *76*, 412–427.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., Hoffman, D., Combrisson, E., Kujala, J., Perrone-Bertolotti, M., et al. (2017). Direct Recordings from Human Anterior Insula Reveal its Leading Role within the Error-Monitoring Network. Cereb. Cortex N. Y. N 1991 *27*, 1545–1557.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. Neuron *47*, 129–141.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. (2008). Associative learning of social value. Nature *456*, 245–249.

Bodi, N., Keri, S., Nagy, H., Moustafa, A., Myers, C.E., Daw, N., Dibo, G., Takats, A., Bereczki, D., and Gluck, M.A. (2009). Reward-learning and the novelty-seeking personality: a between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. Brain *132*, 2385–2395.

Botvinick, M., and Braver, T. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism. Annu. Rev. Psychol. *66*, 83–113.

Boureau, Y.-L., and Dayan, P. (2011). Opponency Revisited: Competition and Cooperation Between Dopamine and Serotonin. Neuropsychopharmacology *36*, 74–97.

Clithero, J.A., and Rangel, A. (2014). Informatic parcellation of the network involved in the computation of subjective value. Soc. Cogn. Affect. Neurosci. *9*, 1289–1302.

Corradi-Dell'Acqua, C., Tusche, A., Vuilleumier, P., and Singer, T. (2016). Cross-modal representations of first-hand and vicarious pain, disgust and fairness in insular and cingulate cortex. Nat. Commun. *7*, 10904.

Craig, A.D., and Craig, A.D. (2009). How do you feel–now? The anterior insula and human awareness. Nat. Rev. Neurosci. *10*.

Deman, P., Bhattacharjee, M., Tadel, F., Job, A.-S., Rivière, D., Cointepas, Y., Kahane, P.,

and David, O. (2018). IntrAnat Electrodes: A Free Database and Visualization Software for Intracranial Electroencephalographic Data Processed for Case and Group Studies. Front. Neuroinformatics *12*.

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage *53*, 1–15.

Fouragnan, E., Retzler, C., and Philiastides, M.G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. Hum. Brain Mapp. *39*, 2887–2906.

Frank, M.J. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. Science *306*, 1940–1943.

Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T., and Hutchison, K.E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. *104*, 16311–16316.

Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. *37*, 1297–1310.

Genovese, C.R., Lazar, N.A., and Nichols, T. (2002). Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. NeuroImage *15*, 870–878.

Haber, S.N., and Knutson, B. (2010). The Reward Circuit: Linking Primate Anatomy and Human Imaging. Neuropsychopharmacology *35*, 4–26.

Khamassi, M., Quilodran, R., Enel, P., Dominey, P.F., and Procyk, E. (2015). Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex. Cereb. Cortex *25*, 3197–3218.

Kim, H., Shimojo, S., and O'Doherty, J.P. (2006). Is Avoiding an Aversive Outcome Rewarding? Neural Substrates of Avoidance Learning in the Human Brain. PLoS Biol. *4*, e233.

Koechlin, E., and Hyafil, A. (2007). Anterior Prefrontal Function and the Limits of Human Decision-Making. Science *318*, 594–598.

Lachaux, J.-P., Chavez, M., and Lutz, A. (2003). A simple measure of correlation across time, frequency and space between continuous brain signals. J. Neurosci. Methods *123*, 175–188.

Lachaux, J.-P., Fonlupt, P., Kahane, P., Minotti, L., Hoffmann, D., Bertrand, O., and Baciu, M. (2007). Relationship between task-related gamma oscillations and BOLD signal: New insights from combined fMRI and intracranial EEG. Hum. Brain Mapp. *28*, 1368–1375.

Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E., and Crone, N.E. (2012). High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research. Prog. Neurobiol. *98*, 279–301.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. (2017).

Behavioural and neural characterization of optimistic reinforcement learning. Nat. Hum. Behav. *1*, 0067.

Liu, X., Hairston, J., Schrier, M., and Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. Neurosci. Biobehav. Rev. *35*, 1219–1236.

Lopez-Persem, A., Bastin, J., Petton, M., Abitbol, R., Lehongre, K., Adam, C., Navarro, V., Rheims, S., Kahane, P., Domenech, P., et al. (2020). Four core properties of the human brain valuation system demonstrated in intracranial signals. Nat. Neurosci.

Manning, J.R., Jacobs, J., Fried, I., and Kahana, M.J. (2009). Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. J. Neurosci. *29*, 13613–13620.

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837–841.

Metereau, E., and Dreher, J.-C. (2013). Cerebral Correlates of Salient Prediction Error for Different Rewards and Punishments. Cereb. Cortex *23*, 477–487.

Miller, E.K., and Cohen, J.D. (2001). An Integrative Theory of Prefrontal Cortex Function. Annu. Rev. Neurosci. *24*, 167–202.

Monosov, I.E., and Hikosaka, O. (2012). Regionally Distinct Processing of Rewards and Punishments by the Primate Ventromedial Prefrontal Cortex. J. Neurosci. *32*, 10318–10330.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science *309*, 951–954.

Niessing, J. (2005). Hemodynamic Signals Correlate Tightly with Synchronized Gamma Oscillations. Science *309*, 948–951.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., and Malach, R. (2007). Coupling between Neuronal Firing Rate, Gamma LFP, and BOLD fMRI Is Related to Interneuronal Correlations. Curr. Biol. *17*, 1275–1285.

Noonan, M.P., Walton, M.E., Behrens, T.E.J., Sallet, J., Buckley, M.J., and Rushworth, M.F.S. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc. Natl. Acad. Sci. *107*, 20547–20552.

Noonan, M.P., Chau, B.K., Rushworth, M.F., and Fellows, L.K. (2017). Contrasting Effects of Medial and Lateral Orbitofrontal Cortex Lesions on Credit Assignment and Decision-Making in Humans. J. Neurosci. *37*, 7023–7035.

O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. Nat. Neurosci. *4*, 95–102.

Oemisch, M., Westendorff, S., Azimi, M., Hassani, S.A., Ardid, S., Tiesinga, P., and Womelsdorf, T. (2019). Feature-specific prediction errors and surprise across macaque fronto-striatal circuits. Nat. Commun. *10*, 176.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. *2011*, 156869.

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., et al. (2012). Critical Roles for Anterior Insula and Dorsal Striatum in Punishment-Based Avoidance Learning. Neuron *76*, 998–1009.

Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. Nat. Commun. *6*, 8096.

Pessiglione, M., and Delgado, M.R. (2015). The good, the bad and the brain: neural correlates of appetitive and aversive values underlying decision making. Curr. Opin. Behav. Sci. *5*, 78–84.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.

Petrovic, P., Pleger, B., Seymour, B., Kloppel, S., De Martino, B., Critchley, H., and Dolan, R.J. (2008). Blocking Central Opiate Function Modulates Hedonic Impact and Anterior Cingulate Response to Rewards and Losses. J. Neurosci. *28*, 10509–10516.

Ramayya, A.G., Pedisich, I., and Kahana, M.J. (2015). Expectation modulates neural representations of valence throughout the human brain. NeuroImage *115*, 214–223.

Rescorla, R.A., and Wagner, A.R. (1972). 3 A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. 18.

Rigoux, L., Stephan, K.E., Friston, K.J., and Daunizeau, J. (2014). Bayesian model selection for group studies — Revisited. NeuroImage *84*, 971–985.

Roesch, M.R., Calu, D.J., Esber, G.R., and Schoenbaum, G. (2010). All That Glitters … Dissociating Attention and Outcome Expectancy From Prediction Errors Signals. J. Neurophysiol. *104*, 587–595.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G.E., and Wager, T.D. (2014). Representation of aversive prediction errors in the human periaqueductal gray. Nat. Neurosci. *17*, 1607–1612.

Rutledge, R.B., Lazzaro, S.C., Lau, B., Myers, C.E., Gluck, M.A., and Glimcher, P.W. (2009). Dopaminergic Drugs Modulate Learning Rates and Perseveration in Parkinson's Patients in a Dynamic Foraging Task. J. Neurosci. *29*, 15104–15114.

Rutledge, R.B., Dean, M., Caplin, A., and Glimcher, P.W. (2010). Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. J. Neurosci. *30*, 13525–13536.

Saez, I., Lin, J., Stolk, A., Chang, E., Parvizi, J., Schalk, G., Knight, R.T., and Hsu, M. (2018). Encoding of Multiple Reward-Related Computations in Transient and Sustained High-Frequency Activity in Human OFC. Curr. Biol.

Schultz, W. (2000). Reward Processing in Primate Orbitofrontal Cortex and Basal Ganglia. Cereb. Cortex *10*, 272–283.

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

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). higher-order learning in humans. *429*, 4.

Seymour, B., O'Doherty, J.P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., and Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat. Neurosci. *8*, 1234–1240.

Sul, J.H., Kim, H., Huh, N., Lee, D., and Jung, M.W. (2010). Distinct Roles of Rodent Orbitofrontal and Medial Prefrontal Cortex in Decision Making. Neuron *66*, 449–460.

Sutton, and Barto (1998). Reinforcement learning: an introduction (Cambridge Univ Press).

Vickery, T.J., Chun, M.M., and Lee, D. (2011). Ubiquity and Specificity of Reinforcement Signals throughout the Human Brain. Neuron *72*, 166–177.

Wallis, J.D. (2012). Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat. Neurosci. *15*, 13–19.

Yacubian, J. (2006). Dissociable Systems for Gain- and Loss-Related Value Predictions and Errors of Prediction in the Human Brain. J. Neurosci. *26*, 9530–9537.

Zaki, J., Wager, T.D., Singer, T., Keysers, C., and Gazzola, V. (2016). The Anatomy of Suffering: Understanding the Relationship between Nociceptive and Empathic Pain. Trends Cogn. Sci. *20*, 249–259.