Neuroscience

Omissions of Threat Trigger Subjective Relief and Reward Prediction Error-Like Signaling in the Human Reward System

Anne L. Willems ✉, **Lukas Van Oudenhove, Bram Vervliet**

Laboratory of Biological Psychology, Department of Brain & Cognition, KU Leuven, Belgium • Leuven Brain Institute, KU Leuven, Belgium • Laboratory for Brain-Gut Axis Studies (LaBGAS), Translational Research in GastroIntestinal Disorders (TARGID), Department of chronic diseases and metabolism, KU Leuven, Belgium

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted
October 16, 2023 (this version)

Posted to bioRxiv
August 17, 2023

Sent for peer review
July 18, 2023

Abstract

The unexpected absence of danger constitutes a pleasurable event that is critical for the learning of safety. Accumulating evidence points to similarities between the processing of absent threat and the well-established reward prediction error (PE). However, clear-cut evidence for this analogy in humans is scarce. In line with recent animal data, we showed that the unexpected omission of (painful) electrical stimulation triggers reward-like activations within key regions of the canonical reward pathway and that these activations correlate with the pleasantness of the reported relief. Furthermore, by parametrically violating participants' probability and intensity related expectations of the upcoming stimulation, we showed for the first time in humans that omission-related activations in the VTA/SN were stronger following omissions of more probable and intense stimulations, like a positive reward PE signal. Together, our findings provide additional support for an overlap in the neural processing of absent danger and rewards in humans.

eLife assessment

This study presents **valuable** findings on the relationship between prediction error and brain activation in response to unexpected omissions of painful electric shock. The strengths are the research question posed, as it has remained unresolved if prediction errors in the context of biologically aversive outcomes resemble reward-based prediction errors. The evidence is **incomplete** due to the task design, which induces a disconnect between verbal instructions and experiential learning, and the lack of analyses accounting for learning and updating of expectations, which are crucial to neural prediction error signaling. The work will be of interest to cognitive neuroscientists and psychologists studying appetitive and aversive learning.

Introduction

The unexpected absence of danger can turn fearful expectations into pleasurable feelings of relief. Realizing that an eerie scraping sound against your bedroom window is not caused by an intruder, but by a tree branch, will probably overwhelm you with a positive hedonic feeling. While such feelings of relief are frequently experienced[1] and readily recognized in others[2], relief is a peculiar emotion. Although it is characterized by a prevailing positive feeling, it is completely dependent on a priori negative expectations: You can only be relieved if you expected/feared the intruder, even though the overall situation is identical as before (no intruder). As it turns out, aversive expectations can transform neutral outcomes into pleasurable events, but why and how this happens in our brain, is not entirely clear.

One situation where the absence of danger is of particular importance is when people learn *not* to be fearful of a previously dangerous situation, a process known as fear extinction. In the lab, these learning situations are mimicked by repeatedly confronting a human or animal with a threat predicting cue (conditional stimulus, CS; e.g. a tone) in the absence of a previously associated aversive event (unconditional stimulus, US; e.g., an electrical stimulation). This results in decreased fear responses to the CS, not by erasing the original fear (CS→US) memory, but rather by generating a new safety (CS→noUS) memory between the CS and the absence of threat[3]. The development of these safety memories is governed by the mismatch (prediction error, PE) between prior expectations of the US and its actual omission, rendering better-than-expected teaching signals that are reminiscent of a reward PE driving reward learning[4–6]. Theoretically, these teaching signals are stronger when the omission was unexpected (e.g. US omissions early in extinction), but progressively decrease as the omission becomes expected (e.g., US omissions late in extinction). Interestingly, subjective experiences of relief follow the same pattern: participants in fear extinction experiments report high levels of relief pleasantness during early US omissions and decreasing relief pleasantness over later omissions[7,8]. In this sense, the emotion of relief might track the PE, when aversive expectations are violated and the outcome is better-than-expected.

Further support that relief is an emotional correlate to omission PE signals comes from our recently developed Expectancy Violation Assessment (EVA) task[9]. In this task, we zoom in on the exact moment of unexpected US omissions outside of a learning context, and we examine how reactions to these omissions are affected by prior expectations. Specifically, participants receive trial-by-trial instructions about the probability (0%, 25%, 50%, 75% and 100%) and intensity (weak, moderate, strong) of a potentially painful upcoming electrical stimulation, time-locked by a countdown clock (see **Fig.1A**). Importantly, most of the trials do not contain the expected stimulation and hence provoke an omission PE (except during 0% trials, where the omission was fully predicted). We found that the experienced pleasantness of the omission-relief reflects the degree of fearful expectation violation, with omissions of more intense and more probable stimulations eliciting more pleasurable feelings of relief, much like a PE-signal.

How the brain processes threat omission PE-signals is not entirely clear yet, but recent fear extinction studies in rodents have revealed a striking similarity with PE processing of unexpected rewards. In the context of reward, it is well-established that dopamine neurons in the ventral tegmental area (VTA) and substantia nigra (SN) of the midbrain increase their firing rate to unexpected rewards (positive PE), suppress their firing for unexpected reward omissions (negative PE) and show no change in firing compared to completely predicted rewards, in line with a formalized PE[10,11]. Like for a positive reward-PE, dopaminergic neurons in the VTA phasically increase their firing rates to early (unexpected), but not late (expected) US omissions during fear extinction[12–16], which consequently triggers downstream dopamine release in the nucleus accumbens (NAc) shell[12,15,16]. Furthermore, optogenetically blocking (or enhancing) the firing rate of these dopaminergic VTA neurons impairs (or facilitates) subsequent fear extinction learning[12–14]. Notably, such dopaminergic VTA/NAc responses to threat

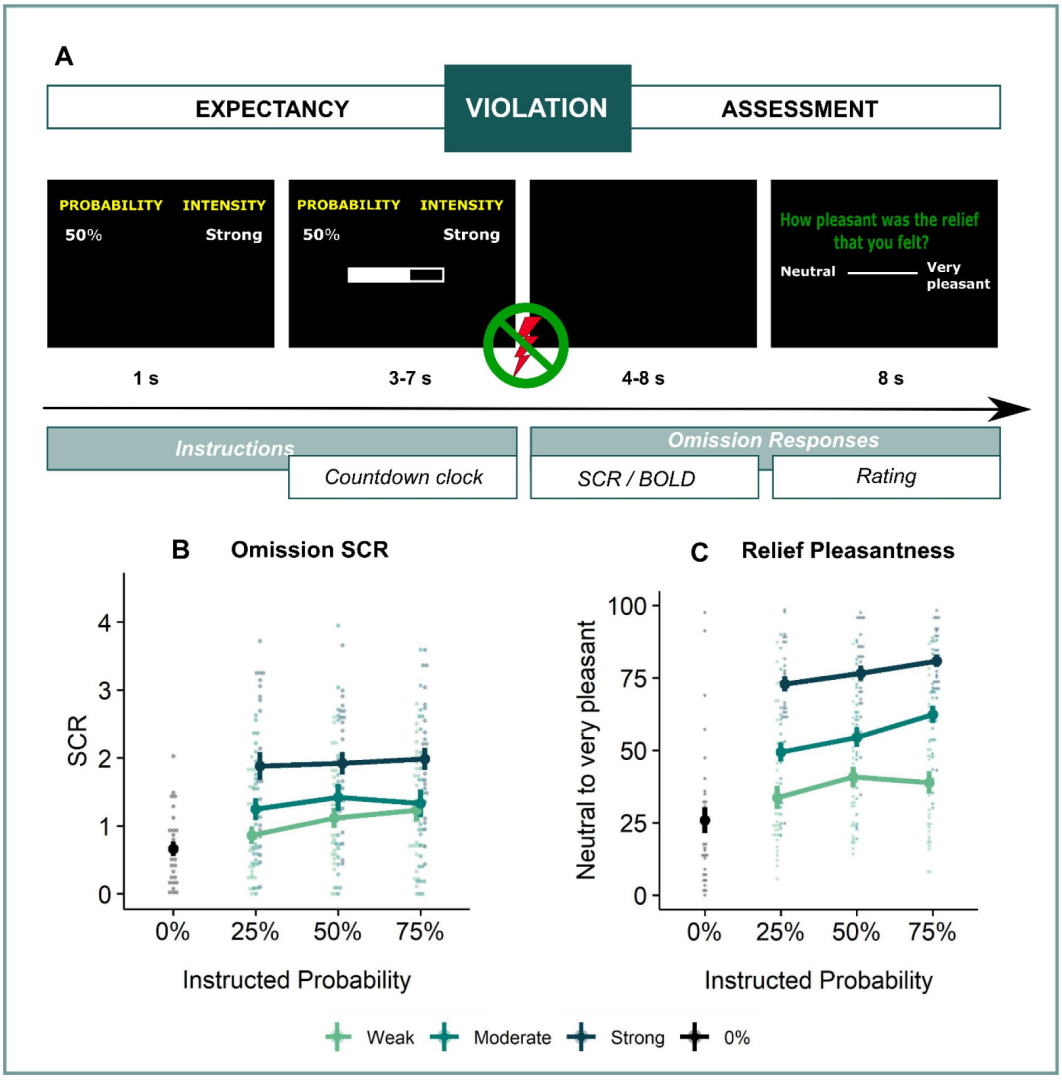

Figure 1.

Experimental Design and Behavioral Results.

A. All trials started with instructions on the probability (ranging from 0% to 100%) and intensity (weak, moderate, strong) of a potentially painful electrical stimulation (1 second), followed by the addition of a countdown bar that indicated the exact moment of stimulation or omission. The duration of the countdown clock was jittered between 3-7 seconds. Then, the screen cleared and the electrical stimulation was either delivered or omitted. Most of the trials (48 trials) did not contain the anticipated electrical stimulation (omission trials). Following a delay of 4 to 8 seconds, a rating scale appeared, probing stimulation-unpleasantness on stimulation trials, and relief-pleasantness on omission trials. After 8 seconds, a random ITI between 4 and 7 seconds started, during which a fixation cross was presented on the screen. The task consisted in total of 72 trials, divided equally over 4 runs (18 trials/run). Each run contained all probability (25, 50, 75) x intensity (weak, moderate, strong) combinations exactly once not followed by the stimulation (9 omission trials), three 0%-omission trials (without any intensity information), three 100%-stimulation trials (followed by the stimulation of the given intensity, one per intensity level per run), and three additional trials from the probability (25,50,75) x Intensity (weak, moderate, strong) matrix that were followed by the stimulation (per run each level of intensity and probability was once paired with the stimulation). For a detailed overview of the trials see Supplementary Figure 1. **B.** SCR were scored as the time integral of the deconvoluted phasic activity (using CDA in Ledalab) within a response window of 1-5s post omission. Responses were square root transformed. SCR were larger following omissions of more probable and more intense US omissions. **C.** The pleasantness of the relief elicited by US omissions was rated on a VAS-scale ranging from 0 (neutral) to 100 (very pleasant). Omission-relief was rated as being more pleasant following omissions of more intense and more probable US. In both graphs, individual data points are presented, with the group averages plotted on top. The error bars represent standard error of the mean.

omissions have also been observed in other experimental tasks, such as conditioned inhibition[17] and avoidance[18,19], confirming that these neural activations match a more general threat omission PE-signal.

In humans, reward-like PE responses to threat omissions during extinction and avoidance have mainly been reported in projection regions of dopaminergic midbrain neurons, such as the ventral striatum (more specifically, NAc and putamen) and prefrontal areas (the ventromedial prefrontal cortex, vmPFC)[20–25]. Activations in these regions correlate with computationally modeled PE signals[20,21,23] and self-reported relief-pleasantness[24], and are modulated by pharmacological manipulation of dopamine receptors[23] and by genetic mutations that are known to enhance striatal phasic dopamine release[20]. Furthermore, connectivity analyses revealed that NAc activations during US omissions in fear extinction were functionally coupled to VTA/SN activations, and that this connectivity was enhanced by the administration of the dopamine precursor L-dopa prior to extinction[23]. The emerging picture is that cortico-striatal activations to unexpected omissions of threat are triggered by dopaminergic inputs from the VTA/SN, just like a positive reward PE, and that these activations play a central role in different types of threat omission-induced learning such as fear extinction and avoidance learning[5,26].

Despite these recent advances, direct observations of threat omission-related VTA/SN responses are currently lacking in humans, which constrains the translation of rodent to human findings regarding the reward PE-like character of threat omission processing. Here, we applied the EVA-task in the MRI scanner to investigate brain responses to unexpected omissions of threat in greater detail, examine their similarity to reward PE-signals, and explore the link with subjective relief. In the context of reward, three necessary and sufficient criteria for PE-signals have been identified (the so-called axiomatic approach[27,28]; which has also been applied to aversive PE[29–31]). A reward PE signal should: (1) be positively related to reward magnitude (larger rewards trigger larger PEs); (2) be negatively related to reward probability (more probable rewards trigger smaller PEs); and (3) not differentiate between fully predicted outcomes of different magnitudes (given that there is no error in prediction, there should be no difference in activation). Thus, in order to assess if threat omission responses in the EVA task qualify as better-than-expected reward PE-signals, we subjected these responses to an axiomatic testing approach. We expected that (1) expected-but-omitted stimulation would trigger increased activity within key areas of the reward circuit (such as the VTA/SN, NAc, left ventral Putamen and vmPFC); that (2) this omission-related activity would fit the three criteria of a positive reward PE, and that (3) this activity would be related to self-reported relief. These hypotheses and the analysis approach were preregistered on Open Science Framework (OSF, *https://osf.io/ugkzf*). Small deviations related to the approach are reported in the supplementary material (Supplementary Note 4).

Results

Self-reported relief and omission SCR track omissions of threat in a PE-like manner

Replicating our previous findings[9], self-reported relief-pleasantness and omission SCR tracked the PE signal during threat omission (see **Fig.1B**/C). Overall, unexpected (non-0%) omissions of threat elicited higher levels of relief-pleasantness and omission SCR than fully expected omissions (0%), evidenced by a main effect of Probability in the 4 (Probability: 0%, 25%, 50%, 75%) x 4 (Run: 1, 2, 3, 4) LMM (For relief pleasantness (N = 31): $F(3,1417) = 188.34$, $p < .001$, $\omega^2 = 0.28$; and for omission SCR (N = 26): $F(3,1190) = 72.90$, $p < .001$, $\omega^2 = 0.15$, with responses to all non-0% probability levels being significantly higher than responses to 0%, p's < .001, Bonferroni-Holm corrected). Furthermore, relief-pleasantness and omission SCR to unexpected omissions (non-0% omissions) increased as a function of instructed Probability and Intensity, in line with the first two PE axioms, evidenced by main effects of Probability (for relief-pleasantness (N = 31): F(2,1031)

=30.64, $p < .001$, $\omega^2 = 0.05$, all corrected pairwise comparisons, p's < .005; for omission SCR (N = 26): F(2,862) = 5.15, $p < .01$, $\omega^2 = 0.01$, with corrected 75% to 25% comparison, $p < .01$), and Intensity (for relief-pleasantness (N = 31): F(2, 1031) = 623.79, $p < .001$, $\omega^2 = 0.55$, all corrected pairwise comparisons, p's < .001; for omission SCR (N = 26): F(2,862.01) = 107.47, $p < .001$, $\omega^2 = 0.20$, all corrected pairwise comparisons, p's < .001) in a 3 (Probability: 25%, 50%, 75%) x 3 (Intensity: weak, moderate, strong) x 4 (Run: 1, 2, 3, 4) LMM. Relief-pleasantness also showed a significant Probability x Intensity interaction (F(4,1031) = 3.76, $p < .005$, $\omega^2 = 0.01$), indicating that the effect of probability was most pronounced for omissions of moderate stimulation (all p's < .05).

Unexpected omissions of threat trigger activations in the VTA/SN and ventral Putamen, but deactivations in the vmPFC

In line with our hypothesis and similar to relief and omission SCR, unexpected (non-0%) omissions of threat elicited on average stronger fMRI activations than fully expected (0%) omissions in the VTA/SN ($t(30) = 4.48$, $p < .001$, $d = 0.81$) and left ventral putamen ($t(30) = 3.50$, $p < .005$, $d = 0.63$) ROIs (see **Fig. 2A**/B). Surprisingly, NAc showed no significant change in activation ($t(30) = -0.59$, $p = .56$) (**Fig. 2D**), and vmPFC showed a significant deactivation ($t(30) = -4.71$, $p < .001$, $d = -0.85$) (**Fig. 2C**). This apparent deactivation could indicate that omission-related responses were lower for unexpected omissions compared to expected omissions. However, it could also have resulted from lingering safety-related vmPFC activations to the 0%-instructions (corresponding to certainty that no stimulation will follow). Such safety-related vmPFC activations are indeed commonly observed during the presentation of CS-in Pavlovian fear conditioning[32]. To exclude this alternative hypothesis, we examined the non0% > 0% contrast during the instruction window. We found no significant difference in vmPFC activation between 0% and non-0% trials, either as ROI average (t(30) = -1.69, $p_{unc} = .1$) or voxel-wise, SVC within the vmPFC mask (see Supplementary Figures 7-9 and Supplementary Tables 3-5 for full description of the anticipatory fMRI activations). This follow-up analysis suggests that the deactivation to unexpected omissions only emerged after the instruction window, and could therefore not be explained by safety-related activation that were obtained during 0% trials.

Omission-related VTA/SN, but not striatal or vmPFC activations increased in a PE-like manner

We next assessed if the omission-related (de)activations could represent reward-like PE-signals by testing the PE axioms for each ROI separately. For axiom 1 and 2, we contrasted omissions following all intensity x probability combination with 0%-omissions and extracted ROI-specific beta averages. These beta-estimates were then entered into linear mixed models that included instructed intensity and probability as regressors of interest, and averaged US-unpleasantness as regressor of no-interest, in addition to a subject-specific intercept. Axiom 3 was tested via a one-sample (two-sided) t-test over the 100%-stimulation versus 0%-omission contrast.

We found that only omission-related **VTA/SN** activations partially fit the profile of a positive reward PE-signal (**Fig. 2A**). Activations were stronger following omissions of more intense (Axiom 1, F(2,240) = 6.14, $p < .005$, $\omega^2 = .04$), and at trend level of more probable threat (Axiom 2, F(2,240) = 2.94, $p = .055$, $\omega^2 = .02$). However, fully predicted stimulations (100%-trials) elicited stronger activations than fully predicted omissions (0%-trials) ($V_{wilcoxon} = 452$, $p < .001$, r = 0.72), contradicting axiom 3 (**Fig 2E**).

Unlike previous findings[20,21,23,24], we found no evidence for striatal reward PE-like processing of threat omissions. While **ventral putamen** responses were stronger following unexpected omissions compared to expected omissions, these activations did not increase with increasing intensity (axiom 1, F(2,240) = 2.29, $p = .10$) or probability (axiom 2, F(2,240) = 0.57, $p =$

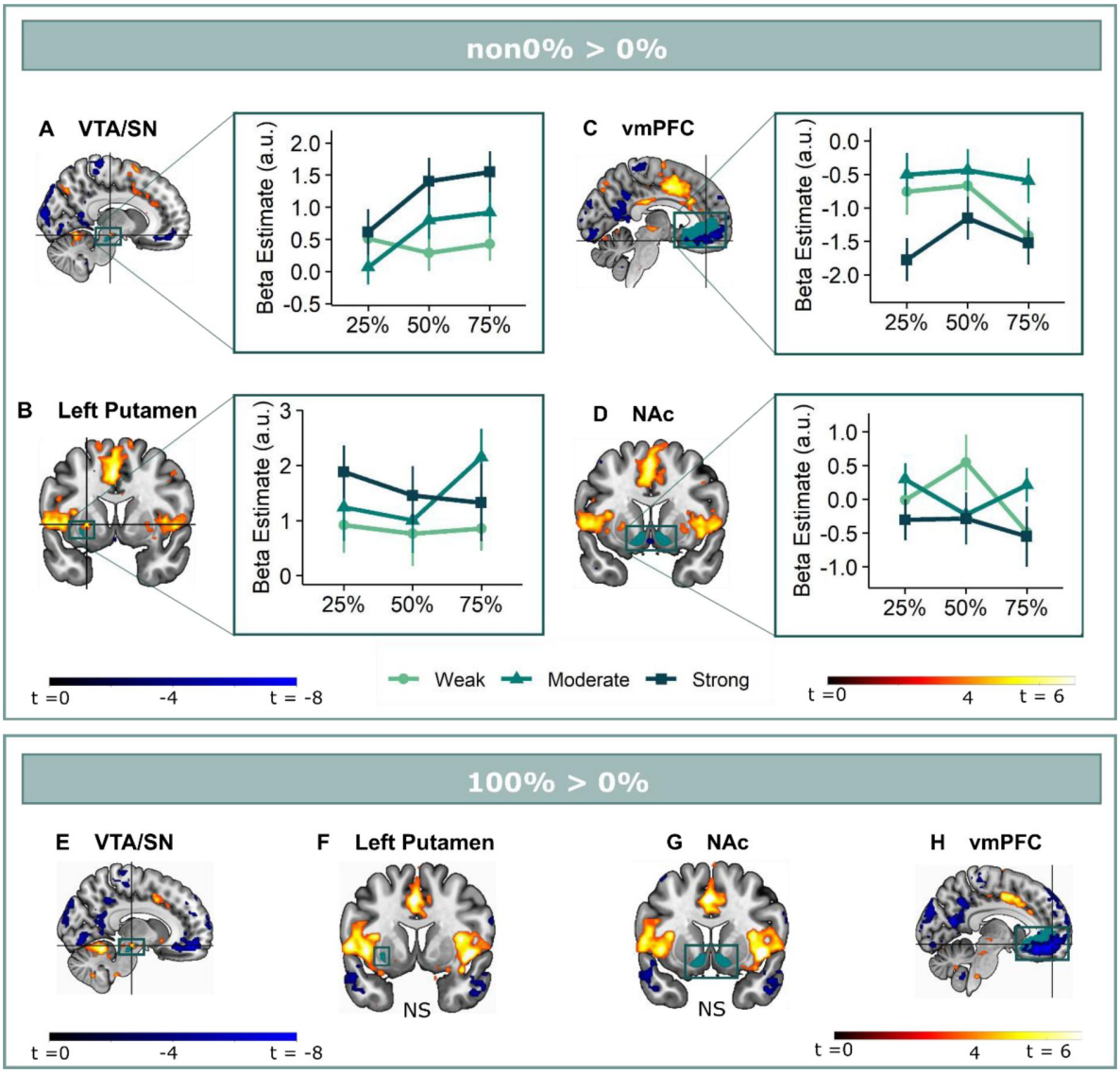

Figure 2.

Omission-related activations in the a priori ROIs.

Unexpected omissions of stimulation (non0% > 0%) triggered significant fMRI responses in **(A)** the VTA/SN, and **(B)** left ventral putamen, but deactivations in **(C)** the vmPFC and no change in activation in **(D)** the NAc. Only for the VTA/SN did the activations increase with increasing Probability and Intensity of omitted stimulation. vmPFC responses decreased with increasing intensity of the omitted stimulation. Fully predicted stimulations elicited stronger activations than fully predicted omission in **(E)** the VTA/SN, no difference in activation for (**F)** the left Putamen and **(G)** the NAc, and stronger deactivations for fully predicted stimulation versus omission in **(H)** the vmPFC. In all figures, the unexpected omission maps were overlayed with the a priori ROI masks (in teal) and were displayed at threshold $p < .001$ (unc) for visualization purposes. The crosshairs represent the peak activation within each a priori ROI. The dots and error bars represent the mean and standard error of the mean.

.57) (**Fig 2B**). Notably, we did find anecdotal evidence that fully predicted stimulations and fully predicted omissions triggered similar activations in the ventral putamen (axiom 3, $t(30) = 1.23$, $p = .46$, BF of 2.62 in favor of the null-hypothesis, **Fig. 2F**). This indicates that ventral putamen activations were exclusively triggered by unexpected threat omissions, and not by fully predicted outcomes, which is similar to a PE-signal.

In general, there was no evidence for omission-related **NAc** activations. Activations were not affected by intensity (axiom 1, $F(2,240) = 1.88$, $p = .16$) or probability instructions (axiom 2, $F(2,240) = 0.75$, $p = .48$) (**Fig 2D**), and while there were no differences in activation between fully predicted stimulation and fully predicted omission ($t(30) = 0.67$, $p = .51$, BF in favor of null hypothesis = 4.24), this equivalence was most likely caused by an overall absence of activation (**Fig 2G**).

Finally, omission-related **vmPFC** deactivations were stronger for omissions of more intense threat (axiom 1, $F(2,240) = 9.29$, $p < .001$, $\omega_p^2 = 0.06$), but were unaffected by probability instructions (axiom 2, $F(2,240) = 1.78$, $p = .17$) (**Fig 2C**). Furthermore, responses were smaller for completely predicted stimulation compared to completely predicted omission (axiom 3, $t(30) = -8.65$, $p < .001$, $d = -1.55$, **Fig 2H**). Taken together, we found no evidence that vmPFC deactivations reflected a positive reward PE-like signal.

A potential explanation for the absent probability effects in the putamen and vmPFC might be that the effects were obscured by including participants who did not believe the probability instructions. Indeed, the provided instructions did not map exactly onto the actually experienced probabilities, but were all followed by stimulation in 25% on the trials. We therefore reran our analyses on a subset of participants who showed probability-related increases in their anticipatory SCR during the countdown clock (N = 21, larger SCR to 75% compared to 25% instructions, see Supplementary Figure 4), which we used as a post-hoc index of actual probability-related expectancy (for detailed analyses of anticipatory SCR, see Supplementary Figure 3). This subgroup analysis revealed no additional effect of Probability for the ventral putamen or the vmPFC, but it rendered the effect of Intensity for the ventral Putamen significant ($F(2,160) = 3.10$ $p < .05$, $\omega^2 = 0.03$). In addition, it increased the effect of probability for the VTA/SN activation ($\omega^2 = .02$ to $.05$), thereby further confirming the PE-profile of omission-related VTA/SN responses.

Anterior insula and dmPFC/aMCC clusters show increased activation for unexpected omissions of threat in a PE-like fashion.

We then explored neural threat omission processing within a wider secondary mask that combined our primary ROIs with additional regions that have previously been associated to reward, pain and PE processing (such as the wider striatum, including the caudate nucleus, and putamen; midbrain nuclei, including the periaqueductal gray (PAG), and red nucleus; medial temporal structures, including the amygdala and hippocampus; midline thalamus, habenula and cortical regions, including the anterior insula (aINS), orbitofrontal (OFC), dorsomedial prefrontal (dmPFC) and anterior cingulate (ACC) cortices). Significant omission-processing clusters (contrast non0% > 0% omissions) were extracted from the mask using a cluster-level threshold ($p < .05$, FWE-corrected), following a primary voxel-threshold ($p < .001$) and included the bilateral anterior insula, bilateral putamen, and a medial cortical cluster encompassing parts of the dorsomedial prefrontal cortex (dmPFC) and the anterior medical cingulate cortex (aMCC) (**Fig. 3A-D**). The left putamen cluster bordered and minimally overlapped with our predefined ventral putamen ROI (4 out of 82 voxels) which was based on the peak PE activity in previous studies[20,21]. Exploratory analyses within a wider whole-brain grey-matter mask identified several other omission-processing clusters (see Supplementary Figure 10, Supplementary Table 6). Probability and Intensity related activity modulations of these clusters can be found in Supplementary Figure 11.

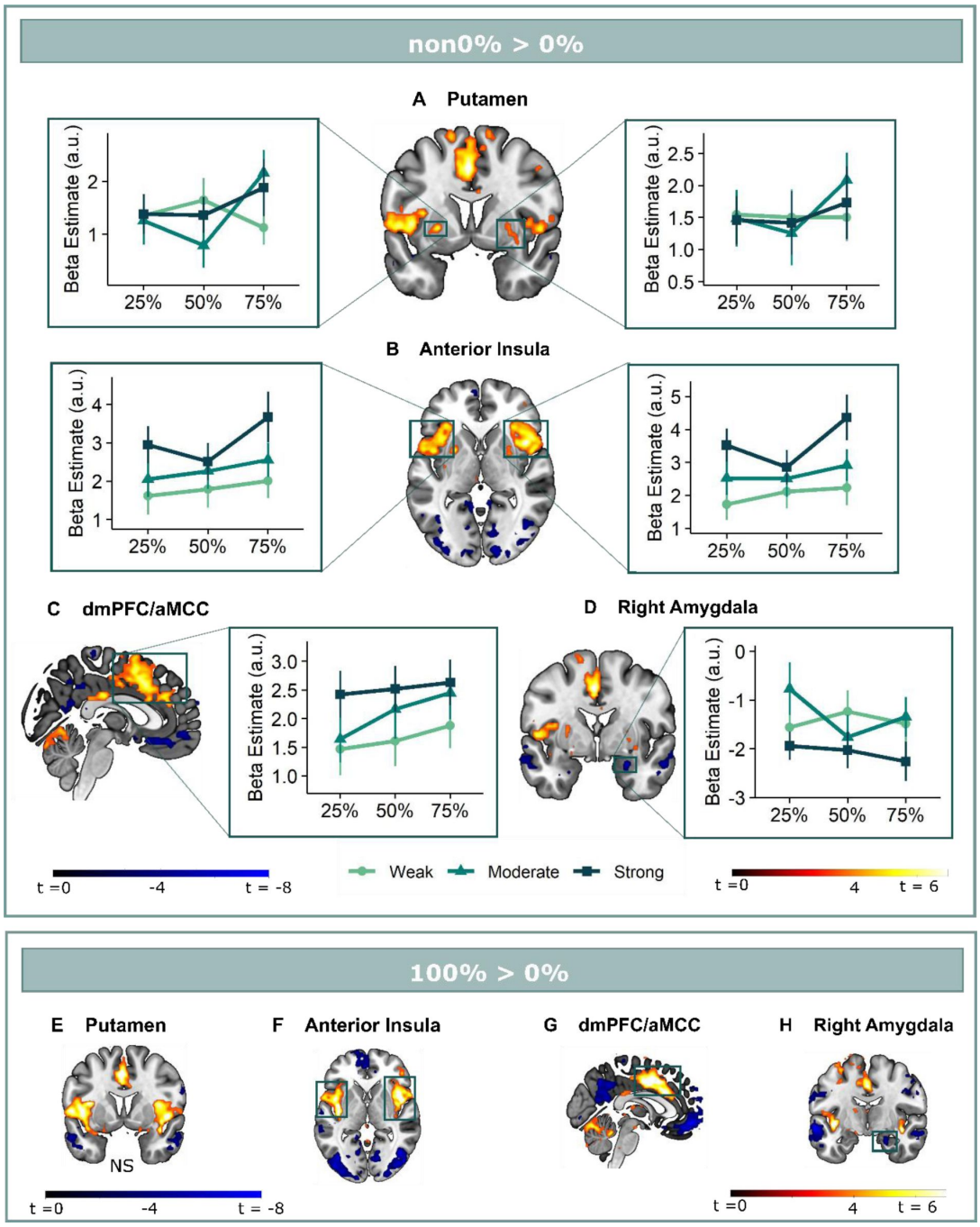

Figure 3.

Omission-related activations in the secondary mask.

We extracted unexpected omission (non0% > 0%) processing clusters within our secondary mask, using a voxel threshold, $p < .001$, followed by a cluster-threshold (FWE-corrected) of $p < .05$. We found significant positive omission processing clusters in **(A)** the bilateral putamen, **(B)** bilateral aINS, **(C)** dmPFC/aMCC, and a trend-level negative omission processing cluster in **(D)** the right amygdala. Omission related activations in the bilateral aINS and dmPFC/aMCC increased with increasing probability and intensity of omitted threat, whereas amygdala activations decreased with omissions of increasingly intense threat. Nevertheless, fully predicted stimulations elicited stronger activations than fully predicted omission in (**F)** the bilateral aIns, and **(G)** the dmPFC/aMCC, and stronger deactivations in the **(H)** right amygdala, but no difference in activation in the **(E)** bilateral putamen. In all figures, the unexpected omission maps are displayed at threshold $p < .001$ (unc) for visualization purposes. The dots and error bars represent the mean and standard error of the mean.

Follow-up analysis of the bilateral **putamen** clusters confirmed that putamen activations did not fit the profile of a positive reward PE. Activations in neither cluster increased with increasing intensity (axiom 1, left: F(2,240) = 0.18, p = .83; right: F(2,240) = 0.06, p = .94) nor probability of threat (axiom 2, left: F(2,240) = 1.41, p = .25; right: F(2,240) = 0.87, p = .42), but like for the a priori ventral Putamen ROI, activations were comparable for fully predicted outcomes, especially in the left cluster (axiom 3, left: V = 339, p = .46, BF = 2.41, right: t(30) = 1.83, p = .46, BF = 1.20).

Positive reward PE-like responses were found in the bilateral **aINS** and **dmPFC/aMCC**, where omission-related activations were stronger following omissions of more intense (left aINS: F(2,240) = 8.95, p < .001, ω^2 = 0.06; right aINS: F(2,240) = 13.49, p < .001, ω^2 = 0.09; dmPFC/aMCC: F(2,240) = 6.59, p < .005, ω^2 = 0.04), and at trend level of more probable threat (left aINS: F(2,240) = 1.87, p =.16, right aINS: F(2,240) = 2.78, p = 0.06, ω^2 = 0.01; dmPFC/aMCC: F(2,240) = 2.48, p =.09, ω^2 = 0.01). Notably aINS and dmPFC/aMCC clusters extended beyond our predefined secondary mask, and including all adjacent above-threshold voxels (p < .001) rendered the effects of probability significant (see Supplementary Figure 11). However, fully predicted stimulations also elicited stronger activations than fully predicted omissions (left aINS: t(30) = 3.21, p < .05, d = 0.58, right aINS: t(30) = 5.26, p < .001, d = 0.95, mdPFC/aMCC: t(30) = 5.57, p < .001, d = 1.00).

Finally, in addition to the vmPFC deactivations (which fell entirely within our vmPFC mask), we found trend-level deactivations for unexpected omission in the **right amygdala** (p = .053). These deactivations were stronger for omissions of more intense (F(2,240) = 3.26, p < .05, ω^2 = 0.02), but not more probable threat (F(2,240) = 0.43, p = .65). Furthermore, fully predicted stimulations triggered larger deactivation than fully predicted omissions (t = -2.77, p = .07, BF = 0.21).

Omission-related activations are related to self-reported relief-pleasantness

We then examined whether omission-related fMRI activations were related to self-reported relief-pleasantness on a trial-by-trial basis. In a pre-registered analysis, we entered z-scored relief-pleasantness ratings as a parametric modulator to the omission regressor in a separate GLM that did not distinguish between the different probability x intensity levels. We found that the VTA/SN (t(30) = 3.26, p < .01, d = 0.59) and ventral putamen ROI (at trend level, t(30) = 2.22, p = .068, d = 0.40) were positively modulated by relief-pleasantness ratings, whereas the vmPFC ROI was negatively modulated by relief-pleasantness ratings (V = 46, p < .001, r = 0.71) (**Fig. 4A-D**). The positive and negative modulations indicate that stronger omission-related activations in the VTA/SN and ventral putamen, and stronger deactivations in the vmPFC were associated with more pleasant relief-reports. Likewise, the bilateral aINS (t > 6.70, p < .001, d > 1.20), dmPFC/aMCC (t(30) = 6.13, p < .001, d = 1.10), and right putamen (t(30) = 4.90, p < .001, d = 0.88), and at trend level left putamen (t(30) = 2.37, p = 0.097, d = 0.43) clusters we identified from the secondary mask were positively modulated, and right amygdala was negatively modulated by relief-pleasantness (V = 59, p < .001, r = 0.67) (**Fig. 4E-H**). Omission-related NAc activation was unrelated to self-reported relief-pleasantness (V = 214, p = 0.52).

A neural signature for relief-pleasantness

The parametric modulation analysis showed that omission-related fMRI activity in our primary and secondary ROIs correlated with the pleasantness of the relief. However, it remains unclear how much of the variance in relief-pleasantness can be predicted by the omission-related fMRI activity pattern across the brain, and what regions contribute most to this prediction. To answer these questions, we trained a LASSO-PCR model (Least Absolute Shrinkage and Selection Operator-Regularized Principle Component Regression) to identify a spatially distributed pattern of brain responses that can predict the perceived pleasantness of the relief (or “neural signature” of

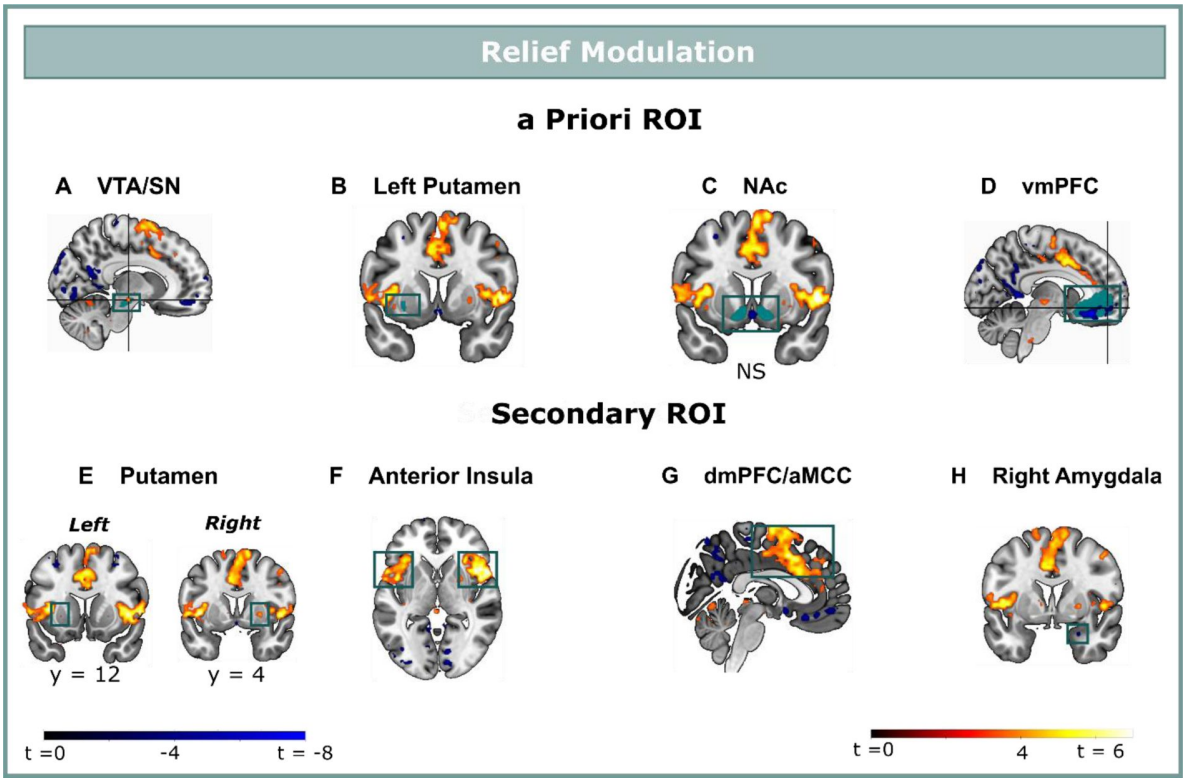

Figure 4.

Relief modulation in the a priori and secondary ROIs.

Omission-related activations were modulated by trial-by-trial levels of relief-pleasantness in **(A)** the VTA/SN, **(B)** left ventral Putamen, and **(D)** vmPFC, but not in **(C)** the NAc. Omission-related activations in the secondary ROIs were modulated by trial-by-trial levels of relief-pleasantness in **(A)** the right putamen, **(B)** bilateral aINS**, (C)** dmPFC**/**aMCC, and the **(D)** right amygdala. In all figures, the relief modulation maps are displayed at threshold $p < .001$ (unc) for visualization purposes.

relief)[33]. We trained the model using fivefold cross-validation with trial-by-trial whole-brain omission-related activation-maps as predictors, and trial-by-trial relief-pleasantness ratings as outcome.

Predicted and reported relief correlated significantly (r = 0.28, p < .001) (**Fig. 5C**), indicating that part of the variance in reported relief-pleasantness could be explained by the neural relief signature response (**Fig. 5A**). Follow-up bootstrap tests (5000 samples) identified a distributed pattern of positive and negative predictive clusters across the brain (**Fig. 5B**, **Table 1**). Increased responses in these clusters predicted increased/decreased relief-pleasantness, respectively. Notably, bootstrap tests indicated that none of our a priori regions of interest significantly contributed to the signature. This was further supported by a pre-registered virtual lesion analysis where we compared the predictive performance of our LASSO-PCR model based on whole-brain data to separate models excluding voxels from our main ROIs in each iteration (see **Fig. 5D**).

Discussion

We examined whether brain reactions to unexpected omissions of threat qualify as positive reward PE signals, and explored their link with subjective relief. We showed that, similar to an unexpected reward, unexpected omissions of stimulation triggered fMRI activations within key regions of the reward pathway (such as the VTA/SN, putamen, dmPFC/aMCC and aINS), and that the magnitude of these activations correlated with the pleasantness of the reported relief. Moreover, omission-related activations in the VTA/SN, the primary reward PE-encoding region in animals[10,11] and humans[34,35], also tracked the probability and intensity of omitted stimulation, in line with the first two criteria of a positive reward-PE signal. In contrast, the NAc and the vmPFC, two other regions that have previously been implicated in reward PE, threat omission processing, and the valuation of rewards and relief[22–24,36–38], showed no (NAc) or even decreased activations (vmPFC) in response to omitted threat; and no correlation (NAc) and a negative correlation (vmPFC) with subjective relief.

Overall, the observed activity pattern of the VTA/SN supports the hypothesis that unexpected omissions of threat are processed as reward PE-like signals in the human brain. However, there are two caveats to this interpretation. First, it remains unclear whether these activations reflect the activity of dopamine cells in this region. The dopamine basis of the reward PE is well established[10,11], and similar VTA dopaminergic responses to threat omissions have been found during fear extinction in rodents[12–14,17]. Yet, the nature of fMRI measurements does not allow us to directly trace back the observed BOLD responses to the phasic firing of dopamine cells at the time of threat omission. Still, the general location of the omission-related VTA/SN activation is consistent with a dopaminergic basis[39,40], as the peak activation falls within a more medial subregion of the SN, which is predominantly composed of dopamine cells (>80% of all cells)[41]. In addition, a recent human fear extinction study found that ingestion of the dopamine precursor L-Dopa increased functional coupling between NAc and VTA/SN at the time of threat omission[23].

A second caveat to the PE-interpretation of VTA/SN activations in the light of the present results is that fully predicted stimulations (100%) triggered stronger activations than fully predicted omissions (0%), which violates the third PE axiom and therefore opposes a PE interpretation. Theoretically, the third axiom states that a pure PE-signal would not differentiate between these fully predicted outcomes, whatever the outcomes are. As such, the violation implies that the VTA/SN responses could not represent PE-signals. Yet, we argue that this axiom should not be a decisive criterion when comparing fully predicted threat to its fully predicted omission. Specifically, a wealth of studies has reported VTA/SN responses to both salient events (even aversive) and PE[10,42], and showed that these responses might be functionally and

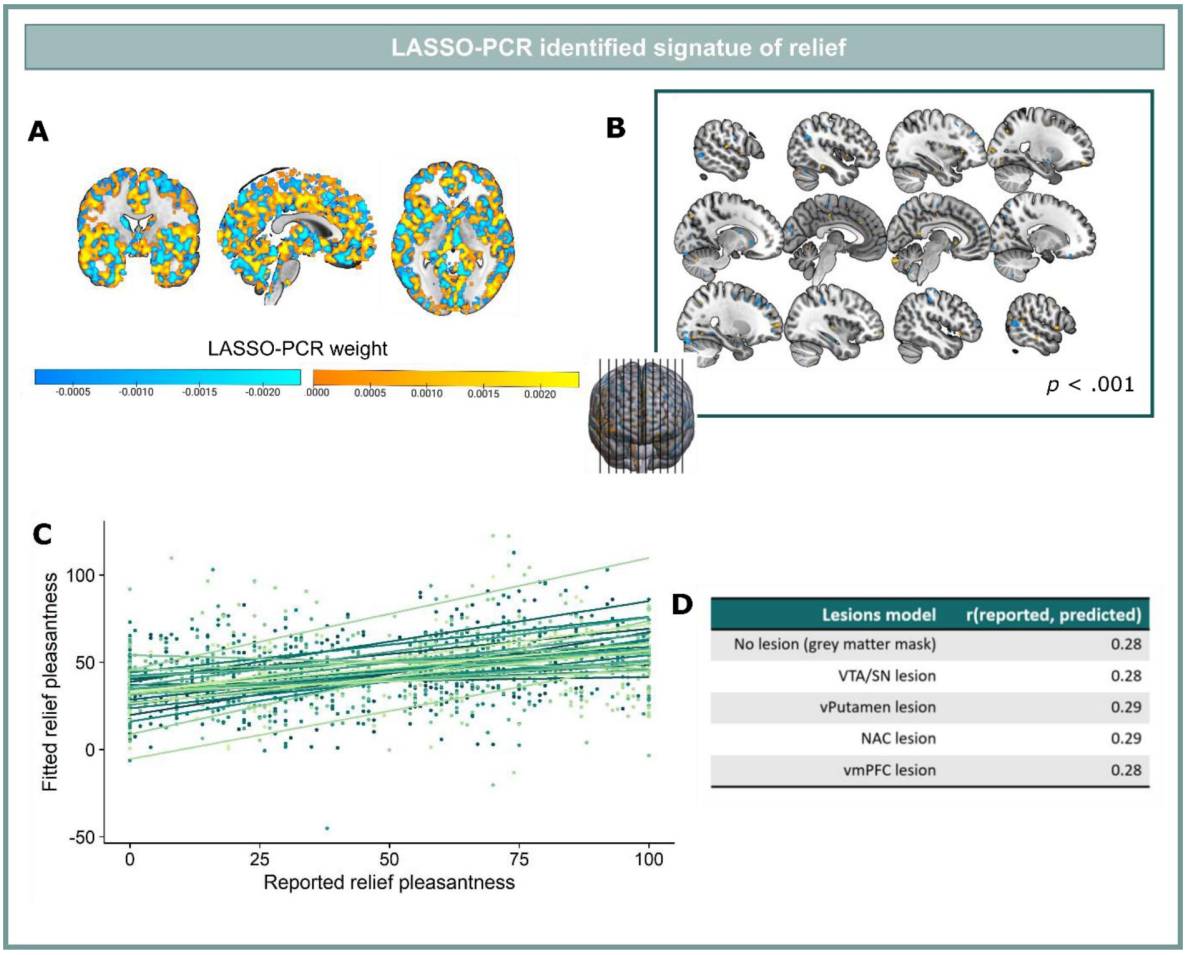

Figure 5.

LASSO-PCR based neural signature of relief-pleasantness.

A. Relief predictive signature map consisting of all voxels within a grey matter mask. All weights were used for prediction. **B.** Thresholded signature map ($p < .001$), consisting of clusters that contribute significantly to the relief prediction (all clusters < 65 voxels). **C.** Predicted and reported relief-pleasantness correlated significantly, $r = 0.28$, $p < .001$. Each line/color represents data of a single participant. **D.** Correlations between reported and predicted relief in all lesion models (in each model one of the ROIs was removed from the grey matter mask). The stable correlation across models confirmed that none of our ROIs contributed significantly to the relief-predictive signature model.

Positive weight clusters						
L/R	*Region*	*K*	*MNI Coordinates (xyz)*			*Z (peak)*
R	Cerebellum Crus 2	17	6	-85	-30	0.006
L	Cerebellum Crus 1	18	-41	-75	-28	0.005
L	Inferior orbitofrontal gyrus	10	-21	14	-23	0.005
R	Inferior temporal gyrus	10	56	-23	-15	0.006
L	Middle orbitofrontal gyrus	15	-25	54	-15	0.007
R	Middle orbitofrontal gyrus	13	38	58	-10	0.007
R	Caudate gyrus	12	4	14	-1	0.007
L	Superior temporal gyrus	11	-55	2	1	0.005
R	Rolandic operculum	19	58	10	3	0.006
R	Superior occipital gyrus	13	20	-103	7	0.006
L	Middle occipital gyrus	14	-23	-97	7	0.006
R	Calcarine gyrus	18	12	-79	10	0.007
L	Middle occipital gyrus	12	-31	-95	14	0.005
R	Middle temporal gyrus	19	54	-47	14	0.005
R	Cuneus	12	4	-79	25	0.006
R	Supramarginal gyrus	12	52	-37	29	0.005
L	Superior occipital gyrus	17	-25	-71	38	0.004
R	Precentral gyrus	15	50	6	43	0.005
L	Superior occipital gyrus	14	-13	-79	43	0.004
L	Mid cingulate gyrus	27	-9	-23	45	0.005
R	Mid cingulate gyrus	12	2	-13	45	0.005
Negative weight clusters						
R	Cerebellum Crus 1	12	30	-69	-28	-0.005
L	Cerebellum Crus 1	11	-17	-89	-19	-0.007
R	Cerebellum Crus 1	41	30	-85	-19	-0.014
R	Fusiform gyrus	10	28	-49	-8	-0.006
R	Superior temporal gyrus	10	52	-2	-6	-0.004
R	Inferior occipital gyrus	11	28	-89	-4	-0.006
R	Superior temporal gyrus	10	66	-17	-4	-0.006
L	Superior temporal gyrus	16	-63	-5	-1	-0.005
L	Caudate	10	-17	18	1	-0.006
L	Middle temporal gyrus	14	-57	-65	-1	-0.004
R	Middle occipital gyrus	12	34	-93	5	-0.005
L	Middle occipital gyrus	11	-39	-89	3	-0.004
R	Middle temporal gyrus	40	58	-57	10	-0.005
L	Calcarine gyrus	18	-11	-81	12	-0.007
L	Cuneus	20	-3	-89	21	-0.006
L	Angular gyrus	16	-45	-59	29	-0.005
L	Middle frontal gyrus	12	-33	46	34	-0.005
R	Postcentral gyrus	64	46	-31	49	-0.005
R	Middle frontal gyrus	19	26	32	45	-0.004
L	Postcentral gyrus	10	-49	-17	45	-0.005
R	Middle frontal gyrus	12	26	10	47	-0.004

Note. Contributing clusters are defined based on voxels-wise FDR-correction with $q < .05$, $k > 10$; *L/R* indicates if the cluster (or peak of the cluster) is part of the left or right hemisphere; *Region* name is identified using the AAL atlas; *K* is the number of voxels in the cluster; *coordinates* are the MNI coordinates of cluster peak; Z is the weight of the cluster peak.

Table 1

Main clusters contributing to the relief signature identified via bootstrapping

anatomically distinct[40,43,44]. Moreover, on a smaller scale, recent electrophysiological studies suggest that dopaminergic PE-signals themselves consist of two components: an initial unselective and short-lasting component that detects any event of sufficient intensity, followed by a subsequent component that codes the prediction error[10]. Thus, given the inherent difference in physical impact (or stimulus salience) of a stimulation versus its omission, the observed increase in activation to stimulations compared to omissions may be expected. Accordingly, our prediction of equal activation to 100%-stimulations and 0%-omissions as a test of the third axiom might have been overoptimistic for the measurement of omission-related PE.

Beyond the midbrain, unexpected omissions of threat also elicited striatal activations that spread across the bilateral putamen (as in[20,21]). Although these activations did not increase with increasing probability and intensity of the omitted stimulation, they were correlated with the reported relief-pleasantness and only occurred when the omission was unexpected (not when the outcome was fully expected, 100%-stimulation or 0% omission). Yet, in contrast to our predictions, the activations did not extend to the NAc. This was surprising, because the NAc is the main striatal projection area of the VTA, and numerous rodent and human studies have attributed reward PE signaling to this region[28,45]. Likewise, two human studies found that self-reported relief and modeled PE-estimates at the time of threat omission covaried with NAc activations[23,24].

Nevertheless, a growing body of research now indicates that human PE encoding is not confined to the NAc, but instead spreads across the striatum. Meta-analyses on reinforcement learning in humans have identified the (left) putamen, and not the NAc as the most consistent reward PE-encoding region across studies[46,47]. Arguably, this emerging functional divergence between rodent and human striatal responses may be linked to neuroanatomical differences at the level of the midbrain. Specifically, where the (medial) VTA and its NAc projections have revealed marked omission-processing in rodents, we found the strongest omission-related activations in the (medial) SN and the putamen, which is a SN projection target. In addition, task-related differences might have contributed to the absent NAc activations. One of the meta-analyses found that the NAc was especially involved in PE-responses when a response is required to obtain the reward/threat omission (e.g., in instrumental tasks)[46]. Arguably, as the EVA task does not allow active control over the omission, it might not have engaged the NAc. Together, our findings call for caution when directly extrapolating rodent findings to human fMRI results.

We found omission-related *deactivations* in the vmPFC that were stronger following omissions of more intense threat, and that correlated with the reported relief. This is again in contrast with our predictions, and with previous studies that showed vmPFC activations during early US omissions in the context of fear extinction[22] and positive associations between vmPFC activity and subjective pleasure[37,38]. Instead, we found vmPFC deactivations for both omissions and stimulations. Interestingly, these deactivations were not limited to the vmPFC, but spanned across key regions of the default mode network[48], such as the PCC and precuneus (see Supplementary Figure 10 and Supplementary Table 6). Furthermore, they were accompanied by widespread activations in key regions of the salience network[49] (such as the VTA/SN, striatum, aINS, dmPFC/aMCC). One potential explanation is therefore that the deactivation resulted from a switch from default mode to salience network, triggered by the salience of the unexpected threat omission or by the salience of the experienced stimulation.

In addition to examining the PE-properties of neural omission responses in our a priori ROIs, we trained a LASSO-PCR model to establish a signature pattern of relief. Interestingly, none of our a priori ROIs appeared to be an important contributor to this neural signature, even though all of them (except NAc) were significantly modulated by relief. Instead, we identified a spatially distributed pattern of brain responses that consisted of several small clusters (all < 65 voxels) across the brain. Some of these clusters fell within other important valuation and error-processing regions in the brain (e.g., OFC, MCC, caudate nucleus). However, all were small (all < 28 voxels) and require further validation in out of sample participants. Still, these data-driven maps suggest

that other regions than our ROIs might have been especially important for the emotional experience of relief and that examining these multivariate patterns can aid our understanding of emotional relief.

Finally, two limitations of the study need to be addressed. First, by aiming to provide a fine-grained analysis of the reward PE-properties of human fMRI responses to threat omission, we focused exclusively on the necessary and sufficient requirements of reward PE signaling, and thereby disregarded another core aspect of PE: its teaching property. In the EVA task expectancies are instructed and all learning is explicitly discouraged. As a result, this task assesses PE completely outside of a learning context. It therefore remains unclear how the PE-signals we observed relate to actual learning. It could for instance be that the observed responses mainly reflected the surprisingness of the outcome, independent of subsequent learning. It therefore remains important to study how the activity patterns of PE-encoding regions are related to expectancy updating in learning paradigms. Furthermore, it would be interesting to see how the neural omission-PE signals are affected in clinical populations. Second, although single unit recordings in rodents have revealed clear PE-like phasic increases in the firing of dopamine cells at the time of threat omission[12–14], one could argue that fMRI measurements cannot capture the same sub-second response. Indeed, it is generally difficult to disentangle prediction *error* responses from the immediately preceding *prediction responses* in fMRI paradigms[50]. Nevertheless, there was no multicollinearity between anticipation and omission regressors in the first-level GLMs (VIFs < 4), making it unlikely that the omission responses purely represented anticipation. Still, because of the slower timescale of fMRI measurements, we cannot conclusively dismiss the alternative interpretation that we assessed (part of) expectancy instead.

In conclusion, by violating instructions about the probability and intensity of a potentially painful stimulation, we found widespread activations in the canonical reward pathway that were furthermore related to PE-like feelings of pleasurable relief. But, more importantly, we showed for the first time in humans that unexpected threat omission triggered VTA/SN activations that partially met the necessary and sufficient criteria of a positive reward PE signal. In doing so, we provided an important missing link for the human translation of the reward-like threat omission processing in rodents.

Methods

Participants

Thirty-one healthy volunteers between the ages of 18 and 25 (mean = 20.65, 19 females) were recruited to participate in our study. All were right-handed and non-smoking, and declared to be free of any serious medical or psychiatric disorder and regular medication use (see Supplementary Notes 1 & 2 and Supplementary Table 1 & 2 for an overview of the exclusion criteria we employed, the sample size rationale and the resulting study sample). Upon their enrollment, participants were furthermore asked to refrain from consuming any alcohol and/or caffeine and exerting any recreational physical exercise in the 24 hours before the scan session. The study was approved by the Ethical committee UZ/KU Leuven (S63852). All participants provided written informed consent and received either partial course credits or a monetary compensation for their participation.

Stimuli

Expectancy Violation Assessment (EVA) Task

The task was an fMRI adaptation of the previously validated EVA task[9] and was programmed in affect5 software[51]. In this task, probability (0%, 25%, 50%, 75%, 100%) and intensity (weak, moderate, strong) information of an upcoming electrical stimulation to the wrist was presented on each trial in the upper left and right corner of the screen respectively. A countdown clock, visualized as a receding bar, indicated the exact moment of stimulation or omission. Responses to the omissions/stimulations were measured.

Electrical Stimulation

The electrical stimulation consisted of a single 2 ms electro-cutaneous pulse, generated by a Digitimer DS8R Bipolar Constant Current Stimulator (Digitimer Ltd, Welwyn Garden City, UK), and delivered via two MR-compatible EL509 electrodes (Biopac Systems, Goleta, CA, USA). Electrodes were filled with Isotonic recording gel (Gel 101; Biopac Systems, Goleta, CA, USA), and were attached to the right wrist. To match the instructions, a total of three intensities were individually selected at the start of the scanning session. The weak stimulus was calibrated to be “mildly uncomfortable”, the moderate stimulus to be “very uncomfortable, but not painful”, and the strong stimulus to be “significantly painful, but tolerable”. Selected Weak (M = 9.35, SD = 4.65), moderate (M = 16.26 mA, SD = 8.13), and strong (M = 34.26, SD = 19.53) differed significantly (Friedman test: $\chi 2(2) = 62$, $p < .001$, W = 1, all Bonferroni-Holm corrected pairwise comparisons, $p < .001$).

Experimental Procedure

Participants were invited to the lab for an intake session, during which exclusion criteria were extensively checked, experimental procedures were explained, and informed consent was obtained. All included participants then filled out the questionnaire battery (see Supplementary Table 2) and were familiarized with the task and rating scales.

At the start of the scanning session, participants were fitted with skin conductance and stimulation electrodes and stimulation intensities were calibrated using a standard workup procedure (for details, see Supplementary Note 3). Participants were then prepared for the scanner and task instructions were repeated. It was emphasized that all trials in the EVA task were independent of one another, meaning that the presence/absence of stimulations on previous trials could not predict the presence/absence of stimulation on future trials. Finally, the selected intensities were presented again and recalibrated if necessary.

In total, the EVA task comprised 72 trials (48 omission trials), divided equally over four runs of 18 trials/run (for a schematic overview of the trial numbers and types, see Supplementary Figure 1). Since we were mainly interested in how omissions of threat are processed, we balanced and maximized these trials by crossing all non-0% probability levels (25, 50, 75, 100) with all intensity levels (weak, moderate, strong) (12 trials). The three 100% trials were always followed by the stimulation of the instructed intensity, while stimulations were omitted in the remaining nine trials. Six additional trials were intermixed in each run: Three 0% trials with the information that no electrical stimulation would follow (akin to 0% Probability information, but without any Intensity information as it does not apply); and three trials from the Probability x Intensity matrix that were followed by electrical stimulation (across the four runs, each Probability x Intensity combination was paired at least once, and at most twice with the electrical stimulation). Within each run trials were presented in a pseudo-random order, with at most two trials with the same intensity or probability as the previous trial.

All trials started with the Probability and Intensity instructions for 1 second, followed by the addition of the countdown bar to the middle of the screen, counting down for 3 to 7 seconds (see **Fig. 1A**). Then, the screen cleared and the electrical stimulation was either delivered or omitted. Following a delay of 4 to 8 seconds (during which skin conductance and BOLD responses to the omission were measured), the rating scale appeared (probing shock-unpleasantness on shock trials and relief-pleasantness on omission trials). The scale remained on the screen for 8 seconds or until the participant responded, followed by a random intertrial interval between 4 and 7 seconds during which only a fixation cross was shown. After the last run, participants were asked some control questions regarding the intensity and probability instructions. Specifically, they were asked how much effort they would exert to prevent future weak/moderate/strong stimulation (from 0 "no effort" to 100 "a lot of effort"); and to estimate how many stimulations they thought they received following each probability instruction.

Subjective ratings

Relief-pleasantness and shock-unpleasantness were probed on omission-and shock-trials using Visual Analogue Scales (VAS) ranging from 0 (neutral) to 100 (very pleasant/unpleasant).

Skin Conductance Responses (SCR)

Fluctuations in skin conductance were measured between two disposable, EL509 electrodes filled with Isotonic recording gel (Gel 101; Biopac Systems, Goleta, CA, USA) that were attached to the hypothenar palm of the left hand. Data were recorded continuously at a 1000 Hz sampling rate via a Biopac MP160 System (Biopac Systems, Goleta, CA, USA)), and Acqknowledge software (version 5.0). Raw data were low-passed filtered at 5 Hz (Butterworth, zerophased) and downsampled to 100 Hz in Matlab (version R2020b), after which they were entered into a continuous decomposition analysis (CDA) with two optimization runs (Ledalab, version 3.4.9[52]). Skin conductance responses (SCR) were scored as the time integral of the deconvoluted phasic activity (integrated SCR) within response windows of 1-4 sec after the onset (anticipatory SCR) and the offset (omission/stimulation SCR) of the countdown clock. Above-threshold responses (>0.01 μS) were square root transformed to reduce the skewness of the distribution. N = 3 participants had incomplete datasets because of a missing run (N = 2) or delayed recording (N = 1), and data of N = 5 were completely excluded for SCR analyses as a result of data loss due to technical difficulties (N = 1) or because they were identified as anticipation non-responder (i.e. participant with smaller average SCR to the clock on 100% than on 0% trials) (N = 4). This resulted in a total sample of N = 26 for all SCR analyses. For all other analyses, the full data set (N = 31) was used (see Supplementary Table 1 for detailed overview).

Statistical Analyses of Ratings and SCR

Rating and SCR data were analyzed in R 4.2.1 (R Core Team, 2022; *https://www.R-project.org*) via linear mixed models that were fit using the lme4 package (v1.1.29[53]). All models included within-subject factors of Intensity and/or Probability and their interaction as fixed effect and a subject-specific intercept as random effect. The intensity factor always contained 3 levels (weak, moderate, strong). The Probability factor depended on the outcome variable. For models of relief and omission SCR that looked at Probability as well as Intensity, Probability had 3 levels (25%, 50%, 75%); but for models of relief and omission SCR that only assessed Probability, Probability had 4 levels (0%, 25%, 50%, 75%). To account for potential changes in the effects of Probability (and Intensity) over time, models included Run (4 levels: 1, 2, 3, 4) and their interaction with Probability (and Intensity) as regressor of no-interest. In addition, we controlled for individual differences in the perceived unpleasantness of the stimulation, by calculating the average of the reported stimulation unpleasantness across the entire task, and entering the resulting (mean-centered) scores as a between-subjects covariate. Inclusion of both regressors of no-interest indeed increased model fit (lower AIC). Model assumptions were checked and influential outliers were identified. Influential outliers were defined as data points above $q_{0.75} + 1.5 * IQR$ or below

$q_{0.25}$ *-1.5*IQR*, with IQR the inter quartile range and $q_{0.25}$ and $q_{0.75}$ corresponding to the first and third quartile, respectively; and with a cook's distance of greater than 4/ number of data points (calculated via influence.ME package in R, v0.9-9[54]). To reduce the influence of these data points, they were rescored to twice the standard deviation from the mean of all data points (corresponding approximately to either .05 and .95 percentile). If results did not change, we report the model including the original data points. If results changed, we report the model with adjusted data points.

Main and interaction effects were evaluated using F-tests and *p*-values that were computed via type III analysis of Variance using Kenward-Rogers degrees of freedom method of the lmerTest package (v3.1.3[55]), and omega squared were reported as an unbiased estimate of effect size (calculated via effectsize package, v0.7.0[56]). All significant effects were followed up with Bonferroni-Holm corrected pairwise comparisons of the estimated marginal means in order to assess the direction of the effect (emmeans package, v1.7.5, Length, 2022). Results related to the regressors of no-interest (Supplementary Figure 5), the anticipatory SCR (Supplementary Figure 3), the post-experimental questions (Supplementary Figure 2), and the stimulation responses (Supplementary Figure 6) are reported in the supplementary material for completeness.

fMRI Analyses

MRI Acquisition

MRI data were acquired on a 3 Tesla Philips Achieva scanner, using a 32-channel head coil, at the Department of Radiology of the University Hospitals Leuven. The four functional runs (226 volumes each) were recorded using an T2*-weighted echo planar imaging sequence (60 axial slices; FOV = 224 x 224 mm; in-plane resolution = 2 x2 mm; interslice gap = 0.2 mm; TR = 2000 ms; TE = 30 ms; MB = 2; flip angle = 90°). In addition, a high resolution T1-weighted anatomical image was acquired for each subject for co-registration and normalization of the EPI data using a MP-RAGE sequence (182 axial slices; FOV = 256 x 240 mm; in-plane resolution = 1 x 1 mm; TR = shortest, TE = 4.6 ms, flip angle = 8°); and a short reverse phase functional run (10 volumes) was acquired using the exact same imaging parameters as the functional runs, but with opposite phase encoding direction. This reverse phase run was used to estimate the B0-nonuniformity map (or fieldmap) to correct for susceptibility distortion. Functional data of one run were missing for N = 4 participants as a result of technical difficulties during scanning. Whenever available, rating and SCR data was still included in the analyses.

fMRI Preprocessing

Prior to preprocessing, image quality was visually checked via quality reports of the anatomical and functional images generated through MRIQC[57]. MRI data were then preprocessed using a standard preprocessing pipeline in fMRIPrep 20.2.3[58]. A detailed overview of the preprocessing steps in fMRPrep can be found in Supplementary Methods 2. In line with our preregistration, spikes were identified and defined as volumes having a framewise displacement (FD) exceeding a threshold of 0.9 mm or DVARS exceeding a threshold of 2. No functional run had more than 15% spikes, and hence none of the runs had to be excluded from our analyses based on our preregistered criterium. Afterwards, the functional data were spatially smoothed with a 4 mm FWHM isotropic Gausian kernel within the "Statistical parametric Mapping" software (SPM12; *https://www.fil.ion.ucl.ac.uk/spm/*).

Subject Level Analysis

Three subject-level general linear models (GLM) were specified. In all models, we concatenated the functional runs and added run-specific intercepts to account for changes over time. The first model investigated the effects of instructed Probability and Intensity on neural omission

processing and therefore included separate stick regressors (duration = 0) for omissions of each Probability x Intensity combination (10 regressors), and stimulations (2 regressors: one for non-100% stimulations, one for 100% stimulations), in addition to boxcar regressors (duration = total duration of event) for the instruction (1 regressor) and rating (1 regressor) windows. The second model assessed how omission fMRI data was related to trial-by-trial relief-pleasantness ratings, by only including a single stick regressor for all omissions (1 regressor) and shocks (1 regressor), and boxcar regressors for instructions (1 regressor) and ratings (1 regressor). Z-scored relief-pleasantness ratings were added as parametric modulator for the omission regressor. A final model estimated single-trial omission responses (48 stick regressors), in addition to a single stick regressor for stimulation and boxcar regressors for instructions (1 regressor) and ratings (1 regressor) and was used for the LASSO-PCR analyses. Regressors in all models were convolved with a canonical hemodynamic response function and a high pass filter of 180 s was applied to remove low frequency drift. Additional non-task related noise was modeled by including nuisance regressors of no-interest for global CSF signal, 24 head motion regressors (consisting of 6 translation and rotation motion parameters and their derivatives, z-scored; and their quadratic terms), and dummy spike regressors.

Group Level Univariate Analysis

Omission processing

To test whether our regions of interest (ROI) were activated by the unexpected omission of threat, we contrasted all non-0% omissions (unexpected omissions) with 0% omissions (expected omissions) at subject-level. Mean activity of each ROI was extracted from the resulting contrast map through marsbar (v0.45[59]), and was entered into group-level (two-sided) one-sample *t*-tests (per ROI) that were Bonferroni-Holm corrected for the total number of ROIs (4 in the main analyses, 10 in the secondary analyses) in R.

We then evaluated whether the observed activity qualified as a 'Prediction Error' by applying an axiomatic testing approach. Specifically, we tested for each ROI if the omission-related activity increased with increasing expected Intensity (axiom 1) and Probability (axiom 2) of threat, and whether completely predicted outcomes (0% omission and 100% stimulations) elicited equivalent activation (axiom 3). For axioms 1 and 2, we extracted mean activity of each ROI from separate omission contrasts that contrasted each omission type (i.e., all possible Probability x Intensity combinations) separately with 0% omissions. These were then entered into a LMM in R including Probability (3 levels: 25%, 50%, 75%) and Intensity (3 levels: weak, moderate, strong) as within-subject factors, and a between-subject covariate of average stimulation unpleasantness (mean-centered) as fixed effects, in addition to a subject-specific intercept. Note that we did not include Run as regressor of no-interest, as Run effects were already accounted for by adding run-specific intercepts to the first level models. Main and interaction effects within each model were followed up with Bonferroni-Holm corrected pairwise comparisons of the estimated marginal means in order to assess the direction of the effect. Finally, to fulfill all necessary and sufficient requirements of a prediction error signal, we contrasted completely predicted omissions (0%) with completely predicted stimulations (100%), as these should trigger equivalent activation in PE-encoding regions. Given that we would expect equivalent activation, Bayes Factors were computed using the BayesFactor package in R (v0.9.12-4.4, Morey & Rouder, 2022) to compare the evidence in favor of alternative and null hypotheses. Larger Bayes Factors indicated more evidence in favor of the null hypotheses.

Parametric modulation of relief

We then tested if the omission-related activity was correlated to self-reported relief-pleasantness on a trial-by-trial basis. To this end, we extracted the mean activity from the modulation contrast for each ROI, and entered these averages in separate one-sample (two-sided) t-test in R, again

correcting for the number of ROIs (4 for main analysis, 10 for secondary analyses). In a subsequent exploratory analysis, we entered z-scored omission SCR-responses as parametric modulator to the omission regressor. Results related to this analysis are reported in Supplementary Figure 13 and Supplementary Table 8.

Neural signature of relief

A LASSO-PCR model (Least Absolute Shrinkage and Selection Operator-Regularized Principal Component Regression) as implemented in CANlab neuroimaging analysis tools (see *https://canlab.github.io/*) was trained using trial-by-trial whole-brain omission-related activation-maps as predictors, and trial-by-trial relief-pleasantness ratings as outcome (for other applications of this approach, see e.g.[33,60–62]). The added value of this machine-learning technique is that relief is predicted across a set of voxels instead of being predicted for each voxel separately (as in standard univariate regression). Specifically, while each voxel in the activation maps is considered a predictor of relief, the LASSO-PCR technique uses a combination of Principle Component Analyses (PCA) and LASSO-regression to (1) group predictive information across individual voxels into larger components (PCA), and (2) maximize the predictive weight of the most informative components by shrinking the regression weight of the least informative components to zero (LASSO regression). Here, we embedded the LASSO-PCR technique within a five-fold cross-validation loop that iteratively trained a LASSO-PCR model in each loop on a different training and validation dataset, which we then averaged across loops to obtain the final model. Model performance was then assessed by calculating the Pearson correlation between reported and model predicted relief. Important features to the signature pattern were identified using bootstrap tests (5000 samples). Furthermore, the contribution of our ROIs to the model's performance was assessed using virtual lesion analysis that consisted of repeating the model training, but excluding voxels within the ROIs and assessing model performance. Note that we also estimated a neural signature of omission SCR, by applying a LASSO-PCR model to predict omission-related SCR responses. Results related to these analyses are reported in Supplementary Figure 14 and Supplementary Table 9.

Regions of Interest (ROI)

Our main regions of interest consisted of key regions of the reward and (relief) valuation pathways such as the Ventral Tegmental Area (VTA) /Substantia Nigra (SN), Nucleus Accumbens (NAC), ventral Putamen, and ventromedial prefrontal cortex (vmPFC). The VTA/SN mask was obtained from Esser and colleagues[23], but was originally defined by Bunzeck and Düzel (2006)[63]. The ventral Putamen ROI was defined as a 6 mm sphere centered around the peak voxel (MNI coordinates: -32,8-6) in the left ventral putamen identified by Raczka et al. (2011)[20], as in Thiele et al. (2021)[21]. However, we overlayed this sphere with a Putamen mask obtained from a high-resolution anatomical atlas for subcortical nuclei defined by Pauli et al. (2018)[64] to assure the mask was restricted to voxels of the putamen and did not extend to the adjacent anterior Insula. Likewise, a bilateral NAc mask was obtained from the Pauli et al. (2018)[64] atlas. The vmPFC mask was obtained by selecting specific parcels from the atlas vmPFC cortex of AFNI (area 14, 32, 24, bilateral). Since we consider the pleasantness of relief from omission of a threat a type of reward, the selection of parcels was made on the base of an activation likelihood estimation (ALE) meta-analysis of 87 studies (1452 subjects) comparing the brain responses to monetary, erotic and food reward outcomes[65].

In addition to our main ROIs, we specified a wider secondary mask that extended our main ROIs with additional regions that have previously been associated to reward, pain and PE processing (such as the wider striatum, including the nucleus caudatus, and putamen; midbrain nuclei, including the periaqueductal gray (PAG), habenula, and red nucleus; limbic structures, including the amygdala and hippocampus; midline thalamus and cortical regions, including the anterior insula (aINS), medial orbitofrontal (OFC), dorsomedial prefrontal (dmPFC) and anterior cingulate (ACC) cortices). Masks for these regions were obtained from CANlab combined atlas (2018) (see

https://canlab.github.io/), and Pauli atlas[64]. Whole brain voxel-wise analyses were restricted to the grey matter mask sparse (from CANlab tools), extended with midbrain voxels (including VTA/SN, RN, habenula). All masks were registered to functional space before analyses and are freely available online at OSF (*https://osf.io/ywpks/*).

Data availability

Raw data files are not freely accessible because one participant did not consent to publish their anonymized data online. Data are available on request by contacting ALW or BV.

Code availability

Analyses code and anatomical masks are freely available at OSF (*https://osf.io/ywpks/*).

Acknowledgements

This research was supported by FWO project grant G078929N (National Research Fund Flanders, Belgium) and C1 project grant C16/19/002 (KU Leuven, Belgium) awarded to BV. LVO is a research professor funded by the KU Leuven Special Research Fund. We would like to thank Mathijs Franssen and Ronald Peeters for their technical support; and Silvia Papalini, Anraoi Rooney, Lieselotte Claes and Anamarija Banjac for their assistance during fMRI data collection.

Author Contribution

AW: Conceptualization, methodology, formal analysis, investigation, data curation, writing – original draft, visualization

LVO: Conceptualization, methodology, formal analysis, writing – review & editing

BV: Conceptualization, methodology, data curation, supervision, writing – original draft, funding acquisition

References

1. Sweeny K., Vohs K. D (2012) **On near misses and completed tasks: The nature of relief** *Psychol. Sci* **23**:464–468

2. Sauter D. A., Scott S. K (2007) **More than one kind of happiness: Can we recognize vocal expressions of different positive states?** *Motiv. Emot* **31**:192–199

3. Bouton M. E., Maren S., McNally G. P (2020) **Behavioral and neurobiological mechanisms of pavlovian and instrumental extinction learning** *Physiol. Rev* :611–681

4. Rescorla R., Wagner A (1972) **A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement** *Class. Cond. Curr. Res. Theory*

5. Papalini S., Beckers T., Vervliet B (2020) **Dopamine: from prediction error to psychotherapy** *Transl. Psychiatry* **10**

6. Salinas-hernández X. I., Duvarci S (2021) **Salinas-hernández, X. I. & Duvarci, S. Dopamine in Fear Extinction. 13, (2021).** *Dopamine in Fear Extinction* **13**

7. Vervliet B., Lange I., Milad M. R (2017) **Temporal dynamics of relief in avoidance conditioning and fear extinction: Experimental validation and clinical relevance** *Behav. Res. Ther* **96**:66–78

8. Papalini S., Ashoori M., Zaman J., Beckers T., Vervliet B (2021) **The role of context in persistent avoidance and the predictive value of relief** *Behav. Res. Ther* **138**

9. Willems A. L., Vervliet B (2021) **When nothing matters : Assessing markers of expectancy violation during omissions of threat** *Behav. Res. Ther* **136**

10. Schultz W (2016) **Dopamine reward prediction-error signalling: A two-component response** *Nat. Rev. Neurosci* **17**:183–195

11. Watabe-Uchida M., Eshel N., Uchida N (2017) **Neural circuitry of reward prediction error** *Annu. Rev. Neurosci* **40**:373–394

12. Luo R., et al. (2018) **A dopaminergic switch for fear to safety transitions** *Nat. Commun* **9**

13. Salinas-Hernández X. I., et al. (2018) **Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes** *Elife* **7**

14. Cai L. X., et al. (2020) **Distinct signals in medial and lateral VTA dopamine neurons modulate fear extinction at different times** *Elife* **9**

15. de Jong J. W., et al. (2019) **A neural circuit mechanism for encoding aversive stimuli in the mesolimbic dopamine system** *Neuron* **101**:133–151

16. Badrinarayan A., et al. (2012) **Aversive stimuli differentially modulate real-time dopamine transmission dynamics within the nucleus accumbens core and shell** *J. Neurosci* **32**:15779–15790

17. Yau J. O. Y., McNally G. P (2018) **Brain mechanisms controlling pavlovian fear conditioning** *J. Exp. Psychol. Anim. Learn. Cogn* **44**:341–357

18. Oleson E. B., Gentry R. N., Chioma V. C., Cheer J. F (2012) **Subsecond dopamine release in the nucleus accumbens predicts conditioned punishment and its successful avoidance** *J. Neurosci* **32**:14804–14808

19. Wenzel J. M., et al. (2018) **Phasic dopamine signals in the nucleus accumbens that cause active avoidance require endocannabinoid mobilization in the midbrain** *Curr. Biol* **28**:1392–1404

20. Raczka K. A., et al. (2011) **Empirical support for an involvement of the mesostriatal dopamine system in human fear extinction** *Transl. Psychiatry* **1**

21. Thiele M., Yuen K. S. L., Gerlicher A. V. M., Kalisch R (2021) **A ventral striatal prediction error signal in human fear extinction learning** *Neuroimage* **229**

22. Lange I., et al. (2020) **Neural responses during extinction learning predict exposure therapy outcome in phobia: results from a randomized-controlled trial** *Neuropsychopharmacology* **45**:534–541

23. Esser R., Korn C. W., Ganzer F., Haaker J (2021) **L-DOPA modulates activity in the vmPFC, nucleus accumbens, and VTA during threat extinction learning in humans** *Elife* **10**

24. Leknes S., Lee M., Berna C., Andersson J., Tracey I (2011) **Relief as a reward: Hedonic and neural responses to safety from pain** *PLoS One* **6**

25. Boeke E. A., Moscarello J. M., LeDoux J. E., Phelps E. A., Hartley C. A (2017) **Active avoidance: Neural mechanisms and attenuation of pavlovian conditioned responding** *J. Neurosci* **37**:4808–4818

26. Kalisch R., Gerlicher A. M. V., Duvarci S (2019) **A dopaminergic basis for fear extinction** *Trends Cogn. Sci* **23**:274–277

27. Caplin A., Dean M (2008) **Axiomatic methods, dopamine and reward prediction error** *Curr. Opin. Neurobiol* **18**:197–202

28. Rutledge R. B., Dean M., Caplin A., Glimcher P. W (2010) **Testing the reward prediction error hypothesis with an axiomatic model** *J. Neurosci* **30**:13525–13536

29. Jepma M., Roy M., Ramlakhan K., van Velzen M., Dahan A (2022) **Different brain systems support learning from received and avoided pain during human pain-avoidance learning** *Elife* **11**

30. Roy M., et al. (2014) **Representation of aversive prediction errors in the human periaqueductal gray** *Nat. Neurosci* **17**:1607–1612

31. Ojala K. E., Tzovara A., Poser B. A., Lutti A., Bach D. R (2022) **Asymmetric representation of aversive prediction errors in Pavlovian threat conditioning** *Neuroimage* **263**

32. Fullana M. A., et al. (2016) **Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies** *Mol. Psychiatry* **21**:500–508

33. Wager T. D., et al. (2013) **An fMRI-based neurologic signature of physical pain** *N. Engl. J. Med* **368**:1388–1397

34. D'Ardenne K., McClure S. M., Nystrom L. E., Cohen J. D (2008) **BOLD responses reflecting dopaminergic signals in the human ventral tegmental area** *Science* **319**:1264–1267

35. Zaghloul K. A., et al. (2009) **Human substantia nigra neurons encode unexpected financial rewards** *Science* **323**:1496–1499

36. Gerlicher A. M. V, Tüscher O., Kalisch R (2018) **Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction** *Nat. Commun* **9**

37. Kim H., Shimojo S., O'Doherty J. P (2006) **Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain** *PLOS Biol* **4**

38. Berridge K. C., Kringelbach M. L (2015) **Pleasure systems in the brain** *Neuron* **86**:646–664

39. Düzel E., et al. (2009) **Functional imaging of the human dopaminergic midbrain** *Trends Neurosci* **32**:321–328

40. Zhang Y., Larcher K. M. H., Misic B., Dagher A (2017) **Anatomical and functional organization of the human substantia Nigra and its connections** *Elife* **6**

41. Root D. H., et al. (2016) **Glutamate neurons are intermixed with midbrain dopamine neurons in nonhuman primates and humans** *Sci. Rep* **6**

42. Diederen K. M. J., Fletcher P. C. (2021) **Diederen, K. M. J. & Fletcher, P. C. Dopamine, prediction error and beyond. Neuroscientist 27, 30–46 (2021).** *Dopamine, prediction error and beyond. Neuroscientist* **27**:30–46

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

44. Pauli W. M., et al. (2015) **Distinct contributions of ventromedial and dorsolateral subregions of the human substantia nigra to appetitive and aversive learning** *J. Neurosci* **35**:14220–14233

45. Hart A. S., Rutledge R. B., Glimcher P. W., Phillips P. E. M (2014) **Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term** *J. Neurosci* **34**:698–704

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

47. Chase H. W., Kumar P., Eickhoff S. B., Dombrovski A. Y (2015) **Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis** *Cogn. Affect. Behav. Neurosci* **15**:435–459

48. Raichle M. E (2015) **The brain's default mode network** *Annu. Rev. Neurosci* **38**:433–447

49. Seeley W. W (2019) **The salience network: A neural system for perceiving and responding to homeostatic demands** *J. Neurosci* **39**:9878–9882

50. Linnman C., Rougemont-Bücking A., Beucke J. C., Zeffiro T. A., Milad M. R (2011) **Unconditioned responses and functional fear networks in human classical conditioning** *Behav. Brain Res* **221**:237–245

51. Spruyt A., Clarysse J., Vansteenwegen D., Baeyens F., Hermans D (2009) **Affect 4.0: A free software package for implementing psychological and psychophysiological experiments** *Exp. Psychol* **57**:36–45

52. Benedek M., Kaernbach C (2010) **A continuous measure of phasic electrodermal activity** *J. Neurosci. Methods* **190**:80–91

53. Bates D., Maechler M., Bolker B., Walker S (2015) **Fitting linear mixed-effects models using lme4** *J. Stat. Softw* **67**:1–48

54. Nieuwenhuis R., te Grotenhuis M, Pelzer B (2012) **Nieuwenhuis, R., te Grotenhuis, M. & Pelzer, B. influence.ME: Tools for detecting influential data in mixed effects models. R J. 4, 38–47 (2012).** *influence.ME: Tools for detecting influential data in mixed effects models. R J* **4**:38–47

55. Kuznetsova A., Brockhoff P. B., C B, H R. (2017) **lmerTEst package: Tests in linear mixed effects models** *J. Stat. Softw* **82**:1–26

56. Ben-Shachar M., Lüdecke D. (2020) **D. effectsize: Estimation of effect size indices and standardized parameters** *J. Open Source Softw* **5**

57. Esteban O., et al. (2017) **MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites** *PLoS One* **12**

58. Esteban O., et al. (2019) **fMRIPrep: a robust preprocessing pipeline for functional MRI** *Nat. Methods* **16**:111–116

59. Brett M., Anton J.-L., Valabregue R., Poline J.-B (2002) **Region of interest analysis using an SPM toolbox** *Presented at the*

60. Wager T. D., Atlas L. Y., Leotti L. A., Rilling J. K (2011) **Predicting individual differences in placebo analgesia: Contributions of brain activity during anticipation and pain experience** *J. Neurosci* **31**:439–452

61. Speer S. P. H., et al. (2023) **A multivariate brain signature for reward** *Neuroimage* **271**

62. Zhou F., et al. (2021) **A distributed fMRI-based signature for the subjective experience of fear** *Nat. Commun* **12**

63. Bunzeck N., Düzel E (2006) **Absolute coding of stimulus novelty in the human substantia nigra/VTA** *Neuron* **51**:369–379

64. Pauli W. M., Nili A. N., Tyszka J. M (2018) **A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei** *Sci. Data* **5**

65. Sescousse G., Caldú X., Segura B., Dreher J. C (2013) **Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies** *Neurosci. Biobehav. Rev* **37**:681–696

Article and author information

Anne L. Willems
Laboratory of Biological Psychology, Department of Brain & Cognition, KU Leuven, Belgium, Leuven Brain Institute, KU Leuven, Belgium
For correspondence: anne.willems@kuleuven.be
ORCID iD: 0000-0003-2234-8916

Lukas Van Oudenhove
Leuven Brain Institute, KU Leuven, Belgium, Laboratory for Brain-Gut Axis Studies (LaBGAS), Translational Research in GastroIntestinal Disorders (TARGID), Department of chronic diseases and metabolism, KU Leuven, Belgium
ORCID iD: 0000-0002-6540-3113

Bram Vervliet
Laboratory of Biological Psychology, Department of Brain & Cognition, KU Leuven, Belgium, Leuven Brain Institute, KU Leuven, Belgium
ORCID iD: 0000-0002-5904-4257

Copyright

Editors

Reviewing Editor
Thorsten Kahnt
National Institute on Drug Abuse Intramural Research Program, United States of America

Senior Editor
Michael Frank
Brown University, United States of America

Reviewer #1 (Public Review):

Summary:
Willems and colleagues test whether unexpected shock omissions are associated with reward-related prediction errors by using an axiomatic approach to investigate brain activation in response to unexpected shock omission. Using an elegant design that parametrically varies shock expectancy through verbal instructions, they see a variety of responses in reward-related networks, only some of which adhere to the axioms necessary for prediction error. In addition, there were associations between omission-related responses and subjective relief. They also use machine learning to predict relief-related pleasantness, and find that none of the a priori "reward" regions were predictive of relief, which is an interesting finding that can be validated and pursued in future work.

Strengths:
The authors pre-registered their approach and the analyses are sound. In particular, the

axiomatic approach tests whether a given region can truly be called a reward prediction error. Although several a priori regions of interest satisfied a subset of axioms, no ROI satisfied all three axioms, and the authors were candid about this. A second strength was their use of machine learning to identify a relief-related classifier. Interestingly, none of the ROIs that have been traditionally implicated in reward prediction error reliably predicted relief, which opens important questions for future research.

Weaknesses:
To ensure that the number of omissions is similar across conditions, the task employs inaccurate verbal instructions; i.e. 25% of shocks are omitted, regardless of whether subjects are told that the probability is 100%, 75%, 50%, 25%, or 0%. Given previous findings on interactions between verbal instruction and experiential learning (Doll et al., 2009; Li et al., 2011; Atlas et al., 2016), it seems problematic a) to treat the instructions as veridical and b) average responses over time. Based on this prior work, it seems reasonable to assume that participants would learn to downweight the instructions over time through learning (particularly in the 100% and 0% cases); this would be the purpose of prediction errors as a teaching signal. The authors do recognize this and perform a subset analysis in the 21 participants who showed parametric increases in anticipatory SCR as a function of instructed shock probability, which strengthened findings in the VTA/SN; however given that one-third of participants (n=10) did not show parametric SCR in response to instructions, it seems like some learning did occur. As prediction error is so important to such learning, a weakness of the paper is that conclusions about prediction error might differ if dynamic learning were taken into account. Lastly, I think that findings in threat-sensitive regions such as the anterior insula and amygdala may not be adequately captured in the title or abstract which strictly refers to the "human reward system"; more nuance would also be warranted.

Reviewer #2 (Public Review):

The question of whether the neural mechanisms for reward and punishment learning are similar has been a constant debate over the last two decades. Numerous studies have shown that the midbrain dopamine neurons respond to both negative and salient stimuli, some of which can't be well accounted for by the classic RL theory (Delgado et al., 2007). Other research even proposed that aversive learning can be viewed as reward learning, by treating the omission of aversive stimuli as a negative PE (Seymour et al., 2004).

Although the current study took an axiomatic approach to search for the PE encoding brain regions, which I like, I have major concerns regarding their experimental design and hence the results they obtained. My biggest concern comes from the false description of their task to the participants. To increase the number of "valid" trials for data analysis, the instructed and actual probabilities were different. Under such a circumstance, testing axiom 2 seems completely artificial. How does the experimenter know that the participants truly believe that the 75% is more probable than, say, the 25% stimulation? The potential confusion of the subjects may explain why the SCR and relief report were rather flat across the instructed probability range, and some of the canonical PE encoding regions showed a rather mixed activity pattern across different probabilities. Also for the post-hoc selection criteria, why pick the larger SCR in the 75% compared to the 25% instructions? How would the results change if other criteria were used?

To test axiom 3, which was to compare the 100% stimulation to the 0% stimulation conditions, how did the actual shock delivery affect the fMRI contrast result? It would be more reasonable if this analysis could control for the shock delivery, which itself could contaminate the fMRI signal, with extra confound that subjects may engage certain behavioral strategies to "prepare for" the aversive outcome in the 100% stimulation condition. Therefore, I agree with the authors that this contrast may not be a good way to test

axiom 3, not only because of the arguments made in the discussion but also the technical complexities involved in the contrast.

Reviewer #3 (Public Review):

Summary:
The authors conducted a human fMRI study investigating the omission of expected electrical shocks with varying probabilities. Participants were informed of the probability of shock and shock intensity trial-by-trial. The time point corresponding to the absence of the expected shock (with varying probability) was framed as a prediction error producing the cognitive state of relief/pleasure for the participant. fMRI activity in the VTA/SN and ventral putamen corresponded to the surprising omission of a high probability shock. Participants' subjective relief at having not been shocked correlated with activity in brain regions typically associated with reward-prediction errors. The overall conclusion of the manuscript was that the absence of an expected aversive outcome in human fMRI looks like a reward-prediction error seen in other studies that use positive outcomes.

Strengths:
Overall, I found this to be a well-written human neuroimaging study investigating an often overlooked question on the role of aversive prediction errors, and how they may differ from reward-related prediction errors. The paper is well-written and the fMRI methods seem mostly rigorous and solid.

Weaknesses:
I did have some confusion over the use of the term "prediction-error" however as it is being used in this task. There is certainly an expectancy violation when participants are told there is a high probability of shock, and it doesn't occur. Yet, there is no relevant learning or updating, and participants are explicitly told that each trial is independent and the outcome (or lack thereof) does not affect the chances of getting the shock on another trial with the same instructed outcome probability. Prediction errors are primarily used in the context of a learning model (reinforcement learning, etc.), but without a need to learn, the utility of that signal is unclear.

An overarching question posed by the researchers is whether relief from not receiving a shock is a reward. They take as neural evidence activity in regions usually associated with reward prediction errors, like the VTA/SN. This seems to be a strong case of reverse inference. The evidence may have been stronger had the authors compared activity to a reward prediction error, for example using a similar task but with reward outcomes. As it stands, the neural evidence that the absence of shock is actually "pleasurable" is limited-albeit there is a subjective report asking subjects if they felt relief.

I have some other comments, and I elaborate on those above comments, below:

1. A major assumption in the paper is that the unexpected absence of danger constitutes a pleasurable event, as stated in the opening sentence of the abstract. This may sometimes be the case, but it is not universal across contexts or people. For instance, for pathological fears, any relief derived from exposure may be short-lived (the dog didn't bite me this time, but that doesn't mean it won't next time or that all dogs are safe). And even if the subjective feeling one gets is temporary relief at that moment when the expected aversive event is not delivered, I believe there is an overall conflation between the concepts of relief and pleasure throughout the manuscript. Overall, the manuscript seems to be framed on the assumption that "aversive expectations can transform neutral outcomes into pleasurable events," but this is situationally dependent and is not a common psychological construct as far as I am aware.

2. The authors allude to this limitation, but I think it is critical. Specifically, the study takes a rather simplistic approach to prediction errors. It treats the instructed probability as the

subjects' expectancy level and treats the prediction error as omission related activity to this instructed probability. There is no modeling, and any dynamic parameters affected by learning are unaccounted for in this design. That is subjects are informed that each trial is independently determined and so there is no learning "the presence/absence of stimulations on previous trials could not predict the presence/absence of stimulation on future trials." Prediction errors are central to learning. It is unclear if the "relief" subjects feel on not getting a shock on a high-probability trial is in any way analogous to a prediction error, because there is no reason to update your representation on future trials if they are all truly independent. The construct validity of the design is in question.

3. Related to the above point, even if subjects veered away from learning by the instruction that each trial is independent, the fact remains that they do not get shocks outside of the 100% probability shock. So learning is occurring, at least for subjects who realize the probability cue is actually a ruse.

4. Bouton has described very well how the absence of expected threat during extinction can create a feeling of ambiguity and uncertainty regarding the signal value of the CS. This in large part explains the contextual dependence of extinction and the "return of fear" that is so prominent even in psychologically healthy participants. The relief people feel when not receiving an expected shock would seem to have little bearing on changing the long-term value of the CS. In any event, the authors do talk about conditioning (CS-US) in the paper, but this is not a typical conditioning study, as there is no learning.

5. In Figure 2 A-D, the omission responses are plotted on trials with varying levels of probability. However, it seems to be missing omission responses in 0% trials in these brain regions. As depicted, it is an incomplete view of activity across the different trial types of increasing threat probability.

6. If I understand Figure 2 panels E-H, these are plotting responses to the shock versus no-shock (when no-shock was expected). It is unclear why this would be especially informative, as it would just be showing activity associated with shocks versus no-shocks. If the goal was to use this as a way to compare positive and negative prediction errors, the shock would induce widespread activity that is not necessarily reflective of a prediction error. It is simply a response to a shock. Comparing activity to shocks delivered after varying levels of probability (e.g., a shock delivered at 25% expectancy, versus 75%, versus 100%) would seem to be a much better test of a prediction error signal than shock versus no-shock.

7. I was unclear what the results in Figure 3 E-H were showing that was unique from panels A-D, or where it was described. The images looked redundant from the images in A-D. I see that they come from different contrasts (non0% > 0%; 100% > 0%), but I was unclear why that was included.

8. As mentioned earlier, there is a tendency to imply that subjects felt relief because there was activity in "the reward pathway."

9. From the methods, it wasn't entirely clear where there is jitter in the course of a trial. This centers on the question of possible collinearity in the task design between the cue and the outcome. The authors note there is "no multicollinearity between anticipation and omission regressors in the first-level GLMs," but how was this quantified? The issue is of course that the activity coded as omission may be from the anticipation of the expected outcome.

10. I did not fully understand what the LASSO-PCR model using relief ratings added. This result was not discussed in much depth, and seems to show a host of clusters throughout the brain contributing positively or negatively to the model. Altogether, I would recommend highlighting what this analysis is uniquely contributing to the interpretation of the findings.