eLife

Reviewed Preprint
v3 • August 23, 2024
Revised by authors

Reviewed Preprint
v2 • April 26, 2024

Reviewed Preprint
v1 • October 16, 2023

Neuroscience

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

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

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 activations within key regions of the reward and salience pathways 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 errors and brain activation in response to unexpected omissions of painful electric shocks. 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 **solid** but there are weaknesses in the experimental design, where verbal instructions do not align with experienced outcome probabilities. It is further unclear how to interpret neural prediction error signaling in the assumed absence of learning. The work will be of interest to cognitive neuroscientists and psychologists studying appetitive and aversive learning.

https://doi.org/10.7554/eLife.91400.3.sa4

Introduction

We experience pleasurable relief when an expected threat stays away[1]. This relief indicates that the outcome we experienced ("nothing") was better than we expected it to be ("threat"). Such a mismatch between expectation and outcome is generally regarded as the trigger for new learning, and is typically formalized as the prediction error (PE) that determines how much there can be learned in any given situation[2]. Over the last two decades, the PE elicited by the absence of expected threat (threat omission PE) has received increasing scientific interest, because it is thought to play a central role in learning of safety. Impaired safety learning is one of the core features of clinical anxiety[3]. A better understanding of how the threat omission PE is processed in the brain may therefore be key to optimizing therapeutic efforts to boost safety learning. Yet, despite its theoretical and clinical importance, research on how the threat omission PE is computed in the brain is only emerging.

To date, the threat omission PE has mainly been studied using fear extinction paradigms that mimic safety learning 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). These (primarily non-human) studies have revealed that there are striking similarities between the PE elicited by unexpected threat omission and the PE elicited by unexpected reward. 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 in response to completely predicted rewards, in line with a formalized PE[4,5]. Likewise, in fear extinction, dopaminergic neurons in the VTA phasically increase their firing rates to early (unexpected), but not late (expected) US omissions [6–10], which consequently triggers downstream dopamine release in the nucleus accumbens (NAc) shell[6,9,10]. Furthermore, optogenetically blocking (or enhancing) the firing rate of these dopaminergic VTA neurons during US omissions impairs (or facilitates) subsequent fear extinction learning[6–8]. Notably, such dopaminergic VTA/NAc responses to threat omissions have also been observed in other experimental tasks, such as conditioned inhibition[11] and avoidance[12,13], 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 ventral putamen) and prefrontal areas (the ventromedial prefrontal cortex, vmPFC)[14–19]. Activations in these regions correlate with computationally modeled PE signals[14,15,17] and are modulated by pharmacological manipulation of dopamine receptors[17] and by genetic mutations that are known to enhance striatal phasic dopamine release[14]. 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[17]. 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[20,21]. Still, direct observations of threat omission-related VTA/SN responses are currently lacking in humans.

As mentioned above, unexpected omissions of threat not only trigger neural activations that resemble a reward PE, they are also accompanied by a pleasurable emotional experience: relief[1]. Because these feelings of relief coincide with the PE at threat omission, relief has been proposed to be an emotional correlate of the threat omission PE[18,22]. Indeed, emerging

evidence has shown that subjective experiences of relief follow the same time-course as theoretical PE during fear extinction. Participants in fear extinction experiments report high levels of relief pleasantness during early US omissions (when the omission was unexpected and the theoretical PE was high) and decreasing relief pleasantness over later omissions (when the omission was expected and the theoretical PE was low)[22,23]. Accordingly, preliminary fMRI evidence has shown that the pleasantness of this relief is correlated to activations in the NAc at the time of threat omission[18]. In that sense, studying relief may offer important insights in the mechanism driving safety learning.

However, is a correlation with the theoretical PE over time sufficient for neural activations/relief to be classified as a PE-signal? In the context of reward, Caplin and colleagues proposed three necessary and sufficient criteria all PE-signals should comply to, independent of the exact operationalizations of expectancy and reward (the so-called axiomatic approach[24,25]; which has also been applied to aversive PE[26–28]). Specifically, the magnitude of a PE signal should: (1) be positively related to the magnitude of the reward (larger rewards trigger larger PEs); (2) be negatively related to likelihood of the reward (more probable rewards trigger smaller PEs); and (3) not differentiate between fully predicted outcomes of different magnitudes (if there is no error in prediction, there should be no difference in the PE signal).

The previously discussed fear conditioning and extinction studies have been invaluable for clarifying the role of the threat omission PE within a learning context[14–17,22,23]. However, these studies were not tailored to create the varying intensity and probability-related conditions that are required to systematically evaluate the threat omission PE in the light of the PE axioms. First, these only included one level of aversive outcome: the electrical stimulation was either delivered or omitted; but the intensity of the stimulation was never experimentally manipulated within the same task. As a result, the magnitude-related axiom could not be tested. Second, as safety learning progressively developed over the course of extinction learning, the most informative trials to evaluate the probability axiom (i.e. the trials with the largest PE) were restricted to the first few CS+ offsets of the extinction phase, and the exact number of these informative trials likely differed across participants as a result of individually varying learning rates. This limited the experimental control and necessary variability to systematically evaluate the probability axiom. Third, because CS-US contingencies changed over the course of the task (e.g. from acquisition to extinction), there was never complete certainty about whether the US would (not) follow. This precluded a direct comparison of fully predicted outcomes. Finally, within a learning context, it remains unclear whether brain responses to the threat omission are in fact responses to the violation of expectancy itself, or whether they are the result of subsequent expectancy updating.

Based on these reasons, we recently developed the Expectancy Violation Assessment (EVA) task[29] in order to study threat omission responses outside of a learning context. By providing verbal instructions on the probability and intensity of an upcoming electrical stimulation, which are then violated by not delivering the stimulation, we showed 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.

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. Specifically, participants received 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**). While stimulations were always delivered on 100% trials and never on 0% trials, most of the other trials (25%-75%) did not contain the expected stimulation and hence provoked an omission PE. 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 5).

Results

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

The verbal instructions were effective at raising the expectation of receiving the electrical stimulation in line with the provided probability and intensity levels. Anticipatory SCR, which we used as a proxy of fearful expectation, increased as a function of the probability and intensity instructions (see Supplemental Figure 3). Accordingly, post-experimental questions revealed that by the end of the experiment participants recollected having received more stimulations after higher probability instructions, and were willing to exert more effort to prevent stronger hypothetical stimulations (see Supplemental Figure 2).

Replicating our previous findings[29], 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). Note that while there was a general drop in reported relief pleasantness and omission SCR over time, the effects of Probability and Intensity remained present until the last run (see Supplementary Figure 5). This further confirms that probability and intensity instructions were effective until the end of the task.

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

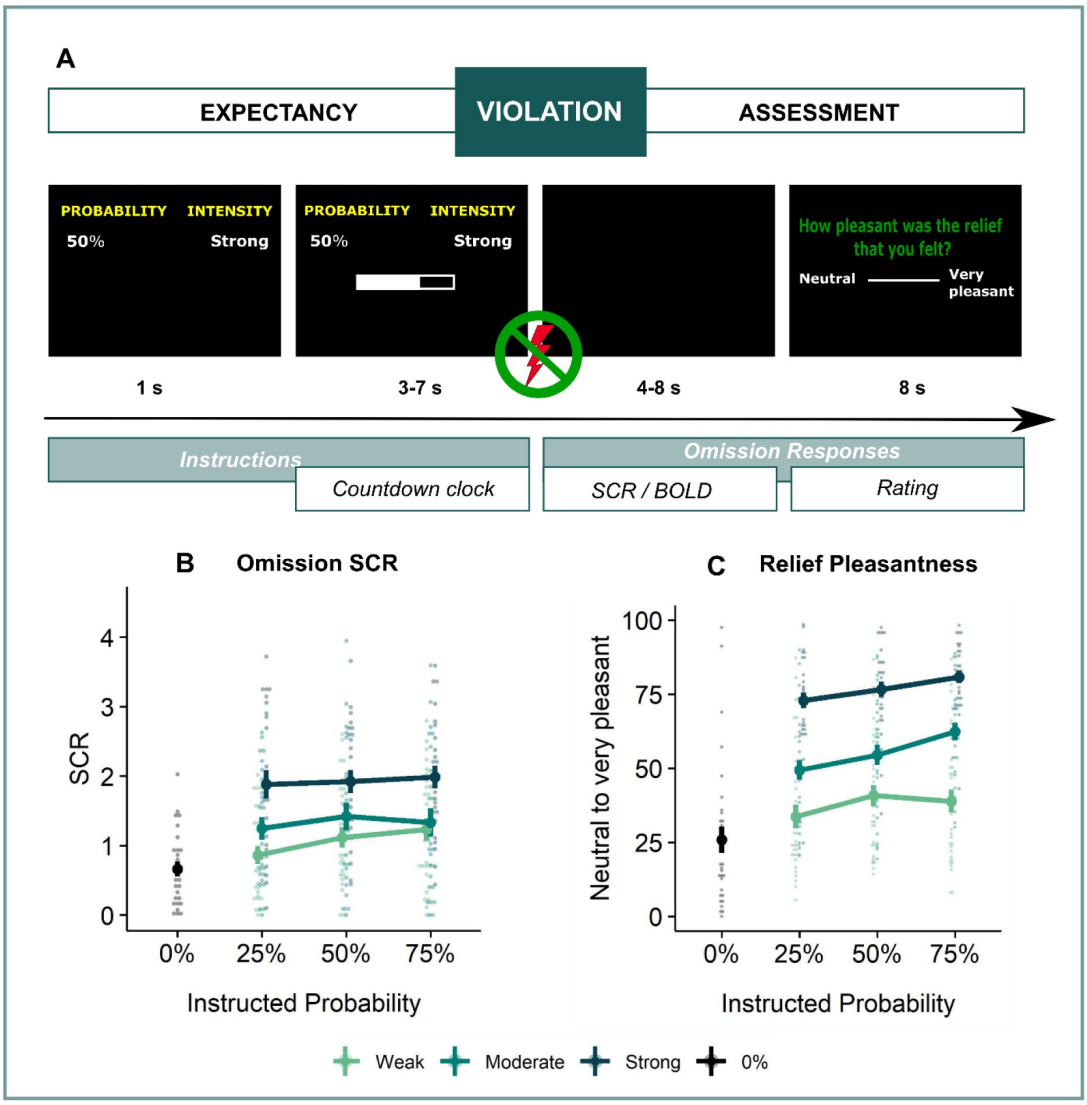

Figure 1.

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, an 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 1-4s 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.

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[30] 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, ω^2 = .04), and at trend level of more probable threat (Axiom 2, F(2,240) = 2.94, p = .055, ω^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[14,15,17,18], 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 = .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, ω^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.

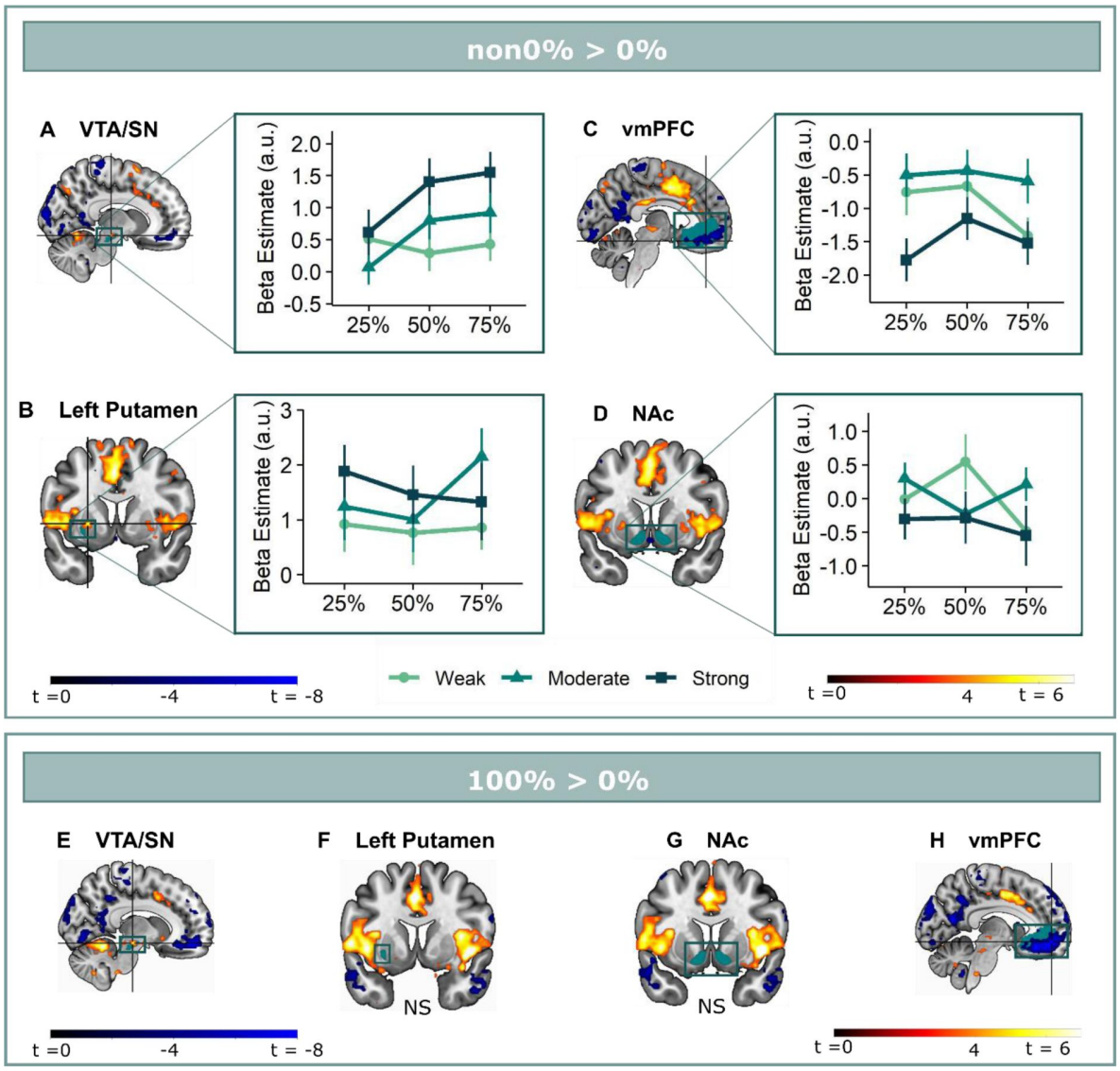

Figure 2.

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 (100%) elicited stronger activations than fully predicted omission (0%) 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 extracted beta-estimates in figures A-D represent the ROI averages from each non-0% > 0% contrast (i.e., 25%>0%; 50%>0%; and 75%>0% for the weak, moderate and strong intensity levels). Any positive beta therefore indicates a stronger activation in the given region compared to a fully predicted omission. Any negative beta indicates a weaker activation. The dots and error bars represent the mean and standard error of the mean.

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 (except for the 0% trials and the 100% 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. 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 $.05$). Likewise, a post-hoc trial-by-trial analysis of the omission-related fMRI activations confirmed that the Probability effect for the VTA/SN activations was stable over the course of the experiment (no Probability x Run interaction) and remained present when accounting for the Gambler's fallacy (i.e., the possibility that participants start to expect a stimulation more when more time has passed since the last stimulation was experienced) (see supplemental note 6). Overall, these post-hoc analyses further confirm 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[14,15]. 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.

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: $F(2,240) = 8.95$, $p < .001$, $\omega^2 = 0.06$; right aINS: $F(2,240) = 13.49$, $p < .001$, $\omega^2 = 0.09$; dmPFC/aMCC: $F(2,240) = 6.59$, $p < .005$, $\omega_p^2 = 0.04$), and at trend level of more probable threat (left aINS: $F(2,240) = 1.87$, $p = .16$, right aINS: $F(2,240) = 2.78$, $p = 0.06$, $\omega_p^2 = 0.01$; dmPFC/aMCC: $F(2,240) = 2.48$, $p = .09$, $\omega_p^2 = 0.01$). 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: $t(30) = 5.57$, $p < .001$, $d = 1.00$).

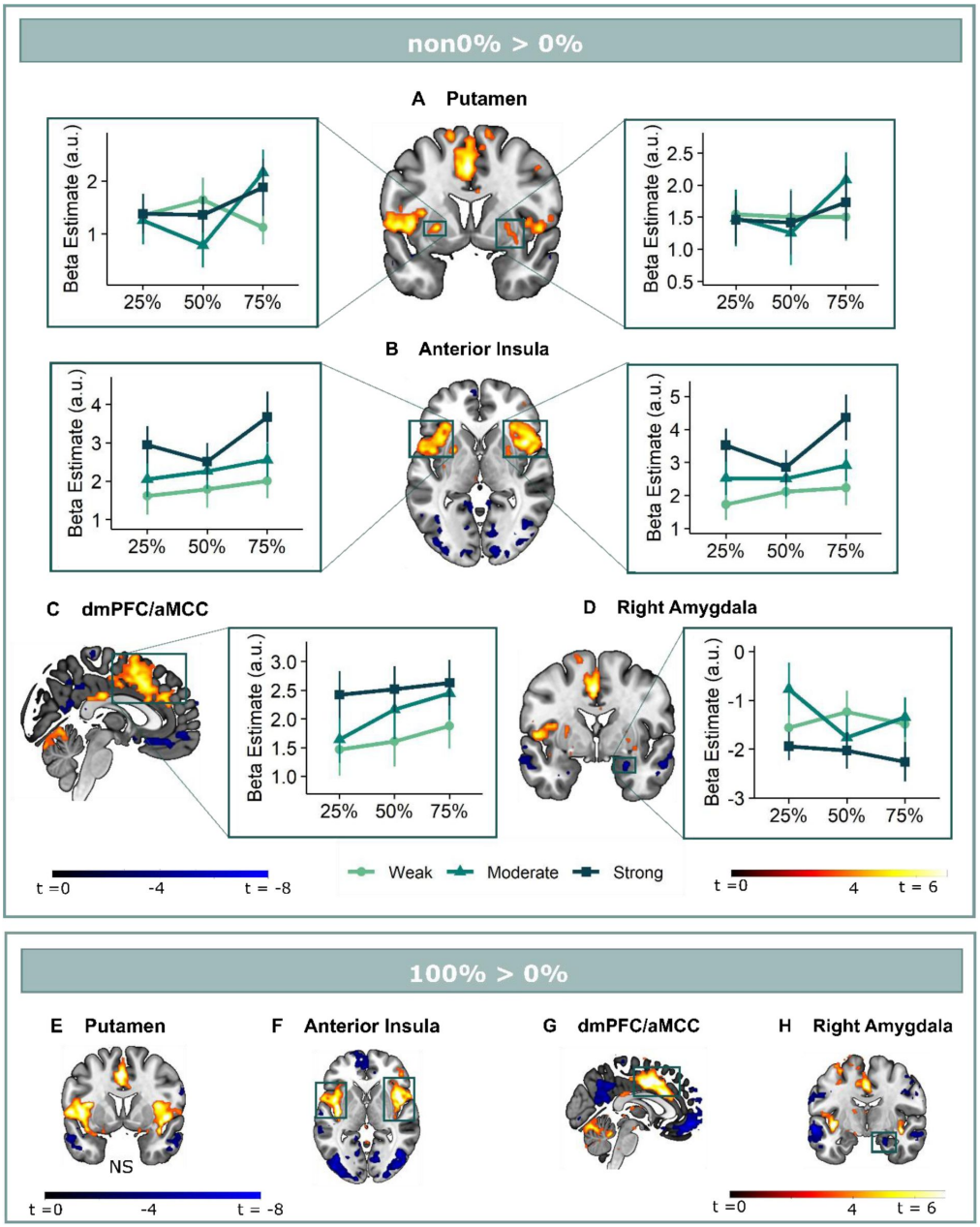

Figure 3.

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 (100%) elicited stronger activations than fully predicted omission (0%) 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 extracted beta-estimates in figures A-D represent the ROI averages from each non-0% > 0% contrast (i.e., 25%>0%; 50%>0%; and 75%>0% for the weak, moderate and strong intensity levels). Any positive beta therefore indicates a stronger activation in the given region compared to a fully predicted omission. Any negative beta indicates a weaker activation. The dots and error bars represent the mean and standard error of the mean.

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 ($F(2,240) = 3.26$, $p < .05$, $\omega^2 = 0.02$), 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 (mass univariate) parametric modulation analysis showed that omission-related fMRI activity in our primary and secondary ROIs correlated with the pleasantness of the relief. However, given that each voxel/ROI is treated independently in this analysis, it remains unclear how the activations were embedded in a wider network of activation across the brain, and which regions contributed most to the prediction of relief. To overcome these limitations, we trained a (multivariate) LASSO-PCR model (Least Absolute Shrinkage and Selection Operator-Regularized Principle Component Regression) in order to identify whether a spatially distributed pattern of brain responses can predict the perceived pleasantness of the relief (or "neural signature" of relief)[31]. Because we used the whole-brain pattern (and not only our a priori ROIs), this analysis is completely data driven and can thus identify which clusters contribute most to the relief prediction. 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**).

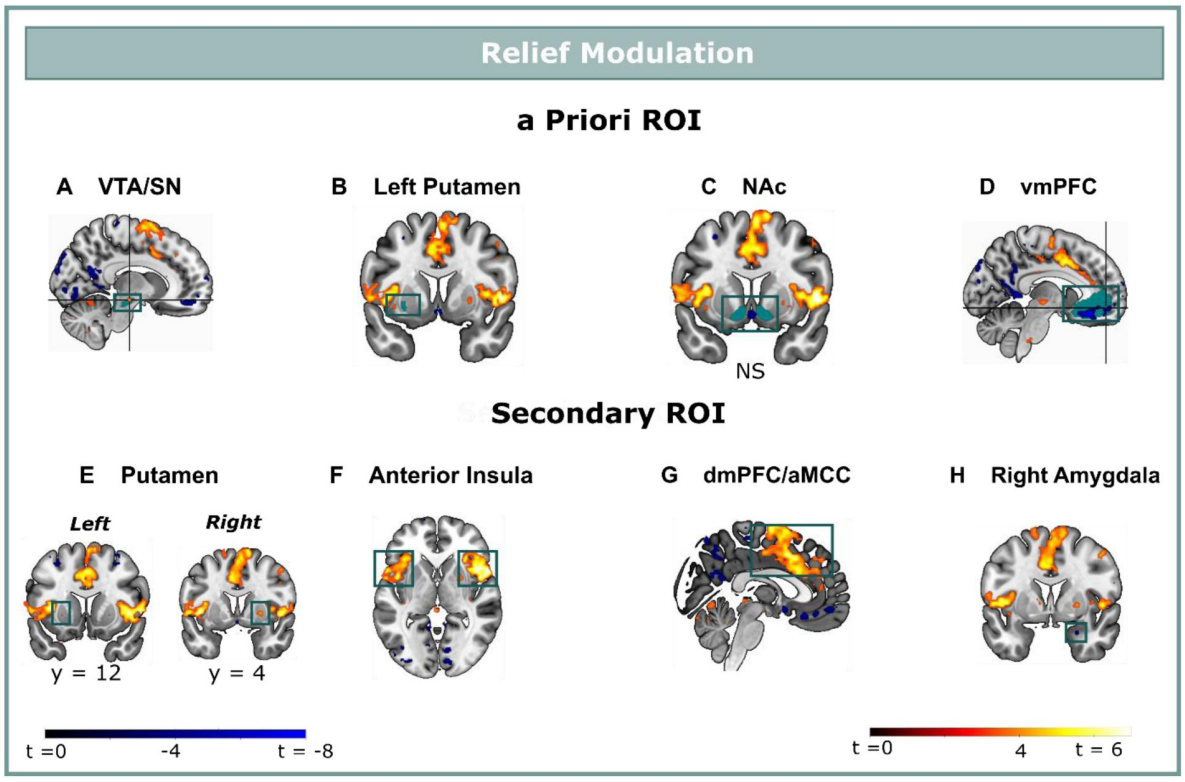

Figure 4.

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.

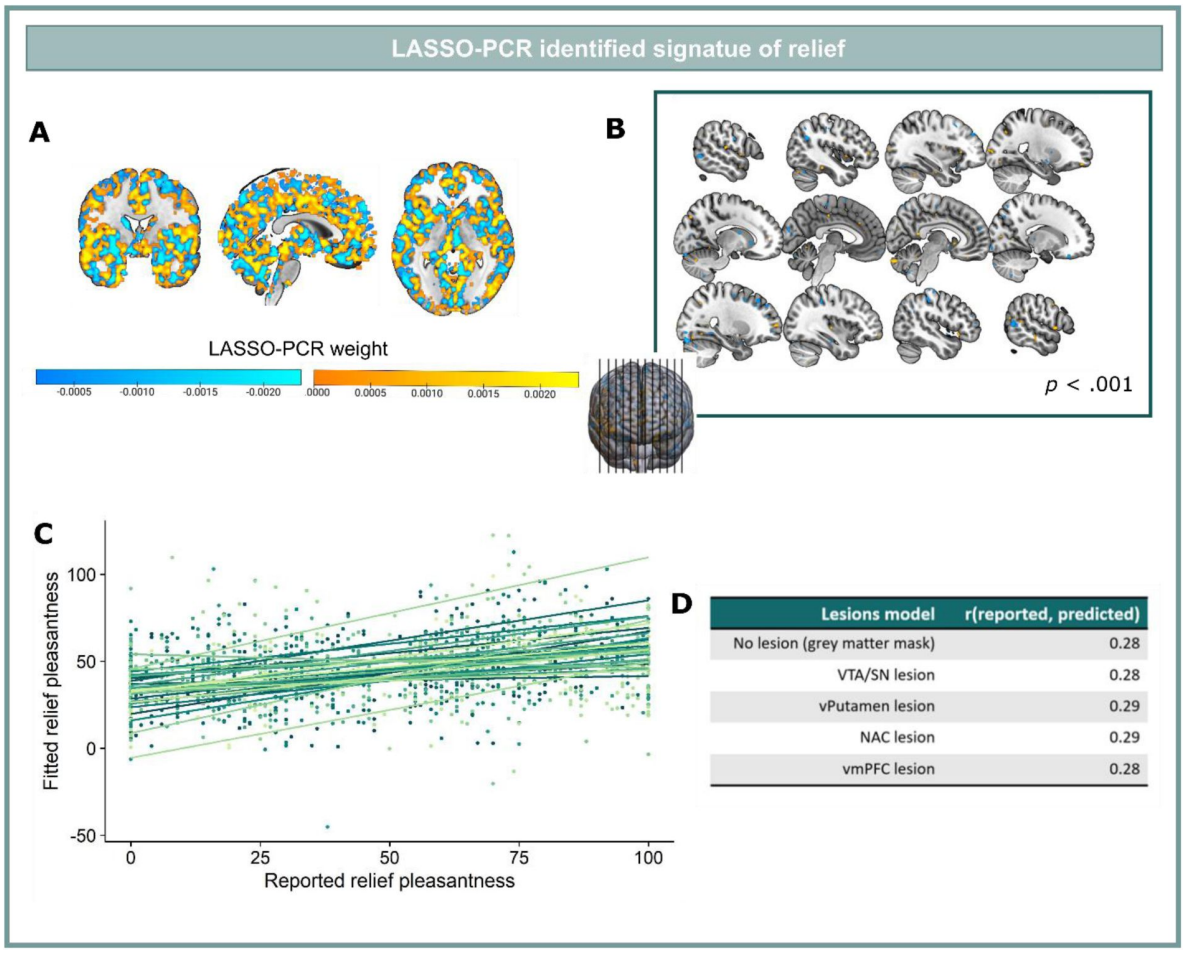

Figure 5.

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.

Table 1

Main clusters contributing to the relief signature identified via bootstrapping

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

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 and salience pathways (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[4,5] and humans[32,33], 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[16–18,34–36], 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[4,5], and similar VTA dopaminergic responses to threat omissions have been found during fear extinction in rodents[6–8,11]. 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[37,38], as the peak activation falls within a more medial subregion of the SN, which is predominantly composed of dopamine cells (>80% of all cells)[39]. 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[17].

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[4,40], and showed that these responses might be functionally and anatomically distinct[38,41,42]. 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[4]. Thus, given that we could not control for the delivery of the stimulation in the 100% > 0% contrast (the delivery of the stimulation completely overlapped with the contrast of interest), it is impossible to disentangle responses to the salience of the stimulation from those to the predictability of the outcome. A fairer evaluation of the third axiom would require outcomes that are roughly similar in terms of salience. When evaluating threat omission PE, this implies comparing fully expected threat omissions following 0% instructions to fully expected absence of stimulation at another point in the task (e.g. during a safe intertrial interval).

Beyond the midbrain, unexpected omissions of threat also elicited striatal activations that spread across the bilateral putamen (as in[14,15]). 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[25,43]. Likewise, two human studies found that self-reported relief and modeled PE-estimates at the time of threat omission covaried with NAc activations[17,18].

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[44,45]. 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)[44]. 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[16] and positive associations between vmPFC activity and subjective pleasure[35,36]. 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[46], 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[47] (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. One interesting finding that only became evident when we compared the univariate and multivariate approach was that none of our a priori ROIs appeared to be an important contributor to the multivariate neural signature, even though all of them (except NAc) were significantly modulated by relief in the univariate analysis. 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[6–8], 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[48]. Nevertheless, there was no multicollinearity between anticipation and omission regressors in the first-level GLMs (Variance Inflation Factor, VIF < 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 reward and salience pathways 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[29] and was programmed in affect5 software[49]. 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: $\chi2(2) = 62$, $p < .001$, 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 wanted to maximize and balance the number of omission trials across the different probability and intensity levels, while also keeping the total number of presentations per probability and intensity instruction constant. Therefore, we crossed 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% omission 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). Note that, based on previous research, we did not expect the inconsistency between the instructed and perceived reinforcement rate to have a negative effect on the Probability manipulation (see Supplementary Note 4). 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 an intertrial interval between 4 and 7 seconds during which only a fixation cross was shown. Note that all phases in the trial were jittered (duration countdown clock, duration outcome window, duration intertrial interval). 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[50]). 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[51]). 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[52]). 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[53]), and omega squared were reported as an unbiased estimate of effect size (calculated via effectsize package, v0.7.0[54]). 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[55]. MRI data were then preprocessed using a standard preprocessing pipeline in fMRIPrep 20.2.3[56]. 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[57]), 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.[31,58–60]). 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[17], but was originally defined by Bunzeck and Düzel (2006)[61]. 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)[14], as in Thiele et al. (2021)[15]. 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)[62] 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)[62] 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[63].

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[62]. 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. ALW was supported by the Internal Funds KU Leuven. 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. Deutsch R., Smith K. J. M., Kordts-Freudinger R., Reichardt R (2015) **How absent negativity relates to affect and motivation: an integrative relief model** *Front. Psychol* **6**

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

3. Beckers T., et al. (2023) **Understanding clinical fear and anxiety through the lens of human fear conditioning** *Nat. Rev. Psychol* **2**:233–245

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

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

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

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

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

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

10. 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

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

12. 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

13. 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

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

15. 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**

16. 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

17. 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**

18. 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**

19. 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

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

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

22. 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

23. 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**

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

25. 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

26. 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**

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

28. 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**

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

30. 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

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

32. 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

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

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

35. 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**

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

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

38. 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**

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

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

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

42. 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

43. 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

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

45. 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

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

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

48. 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

49. 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

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

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

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

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

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

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

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

57. Brett M., Anton J.-L., Valabregue R., Poline J.-B (2002) **Region of interest analysis using an SPM toolbox** *8th International Conference on Functional Mapping of the Human Brain*

58. 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

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

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

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

62. 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**

63. 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

Editors

Reviewing Editor

Thorsten Kahnt

National Institute on Drug Abuse Intramural Research Program, Baltimore, United States of America

Senior Editor
Michael Frank
Brown University, Providence, 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:

The authors have done many analyses to address weaknesses in response to reviews. I will still note that 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 using quantitative models.

https://doi.org/10.7554/eLife.91400.3.sa3

Reviewer #2 (Public Review):

The paper by Willems aimed to uncover the neural implementations of threat omissions and showed that the VTA/SN activity was stronger following the omissions of more probable and intense shock stimulation, mimicking a reward PE signal.

My main concern remains regarding the interpretation of the task as a learning paradigm (extinction) or simply an expectation violation task (difference between instructed and experienced probability), though I appreciate some of the extra analyses in the responses to the reviewers. Looking at both the behavioral and neural data, a clear difference emerges among different US intensities and non-0% vs. 0% contrasts, however, the difference across probabilities was not clear in the figures, potentially partly due to the false instructions subjects received about the shock probabilities.

The lack of probability related PE demonstration, both in behavior and to a less extent in imaging data, does not fully support the PE axioms (0% and 100% are by themselves interesting categories since the instruction and experience matched well and therefore might need to be interpreted differently from other probabilities).

As the other reviewers pointed out, the application of instruction together with extinction paradigm complicates the interpretation of results. Also, the trial-by-trial analysis suggestion was responded by the probability x run interaction analysis, which still averaged over trials within each run to estimate a beta coefficient. So my evaluation remains that this is a valuable study to test PE axioms in the human reward and salience systems but authors need to be extremely careful with their wordings as to why this task is not a particularly learning paradigm (or the learning component did not affect their results, which was in conflict with the probability related SCR, pleasantness ratings as well as BOLD signals).

https://doi.org/10.7554/eLife.91400.3.sa2

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.

Once again, the authors were very responsive to feedback. I have no further comments.

https://doi.org/10.7554/eLife.91400.3.sa1

Author response:

The following is the authors' response to the previous reviews.

> ***Reviewer #1 (Public Review):***

The reviewer retained most of their comments from the previous reviewing round. In order to meet these comments and to further examine the dynamic nature of threat omission-related fMRI responses, we now re-analyzed our fMRI results using the single trial estimates. The results of these additional analyses are added below in our response to the recommendations for the authors of reviewer 1. However, we do want to reiterate that there was a factually incorrect statement concerning our design in the reviewer's initial comments. Specifically, the reviewer wrote that "25% of shocks are omitted, regardless of whether subjects are told that the probability is 100%, 75%, 50%, 25%, or 0%." We want to repeat that this is not what we did. 100% trials were always reinforced (100% reinforcement rate); 0% trials were never reinforced (0% reinforcement rate). For all other instructed probability levels (25%, 50%, 75%), the stimulation was delivered in 25% of the trials (25% reinforcement

rate). We have elaborated on this misconception in our previous letter and have added this information more explicitly in the previous revision of the manuscript (e.g., lines 125-129; 223-224; 486-492).

> ***Reviewer #1 (Recommendations For The Authors):***
>
> *I do not have any further recommendations, although I believe an analysis of learning-related changes is still possible with the trial-wise estimates from unreinforced trials. The authors' response does not clarify whether they tested for interactions with run, and thus the fact that there are main effects does not preclude learning. I kept my original comments regarding limitations, with the exception of the suggestion to modify the title.*

We thank the reviewer for this recommendation. In line with their suggestion, we have now reanalyzed our main ROI results using the trial-by-trial estimates we obtained from the firstlevel omission>baseline contrasts. Specifically, we extracted beta-estimates from each ROI and entered them into the same Probability x Intensity x Run LMM we used for the relief and SCR analyses. Results from these analyses (in the full sample) were similar to our main results. For the VTA/SN model, we found main effects of Probability ($F = 3.12$, $p = .04$), and Intensity ($F = 7.15$, $p < .001$) (in the model where influential outliers were rescored to 2SD from mean). There was no main effect of Run ($F = 0.92$, $p = .43$) and no Probability x Run interaction ($F = 1.24$, $p = .28$). If the experienced contingency would have interfered with the instructions, there should have been a Probability x Run interaction (with the effect of Probability only being present in the first runs). Since we did not observe such an interaction, our results indicate that even though some learning might still have taken place, the main effect of Probability remained present throughout the task.

There is an important side note regarding these analyses: For the first level GLM estimation, we concatenated the functional runs and accounted for baseline differences between runs by adding run-specific intercepts as regressors of no-interest. Hence, any potential main effect of run was likely modeled out at first level. This might explain why, in contrast to the rating and SCR results (see Supplemental Figure 5), we found no main effect of Run. Nevertheless, interaction effects should not be affected by including these run-specific intercepts.

Note that when we ran the single-trial analysis for the ventral putamen ROI, the effect of intensity became significant ($F = 3.89$, $p = .02$). Results neither changed for the NAc, nor the vmPFC ROIs.

> ***Reviewer #2 (Public Review):***
>
> *Comments on revised version:*
>
> *I want to thank the authors for their thorough and comprehensive work in revising this manuscript. I agree with the authors that learning paradigms might not be a necessity when it comes to study the PE signals, but I don't particularly agree with some of the responses in the rebuttal letter ("Furthermore, conditioning paradigms generally only include one level of aversive outcome: the electrical stimulation is either delivered or omitted."). This is of course correct description for the conditioning paradigm, but the same can be said for an instructed design: the aversive outcome was either delivered or not. That being said, adopting the instructed design itself is legitimate in my opinion.*

We thank the reviewer for this comment. We have now modified the phrasing of this argument to clarify our reasoning (see lines 102-104: “First, these only included one level of aversive outcome: the electrical stimulation was either delivered at a fixed intensity, or omitted; but the intensity of the stimulation was never experimentally manipulated within the same task.”).

The reason why we mentioned that “the aversive outcome is either delivered or omitted” is because in most contemporary conditioning paradigms only one level of aversive US is used. In these cases, it is therefore not possible to investigate the effect of US Intensity. In our paradigm, we included multiple levels of aversive US, allowing us to assess how the level of aversiveness influences threat omission responding. It is indeed true that each level was delivered or not. However, our data clearly (and robustly across experiments, see Willems & Vervliet, 2021) demonstrate that the effects of the instructed and perceived unpleasantness of the US (as operationalized by the mean reported US unpleasantness during the task) on the reported relief and the omission fMRI responses are stronger than the effect of instructed probability.

> *My main concern, which the authors spent quite some length in the rebuttal letter to address, still remains about the validity for different instructed probabilities. Although subjects were told that the trials were independent, the big difference between 75% and 25% would more than likely confuse the subjects, especially given that most of us would fall prey to the Gambler's fallacy (or the law of small numbers) to some degree. When the instruction and subjective experience collides, some form of inference or learning must have occurred, making the otherwise straightforward analysis more complex. Therefore, I believe that a more rigorous/quantitative learning modeling work can dramatically improve the validity of the results. Of course, I also realize how much extra work is needed to append the computational part but without it there is always a theoretical loophole in the current experimental design.*

We agree with the reviewer that some learning may have occurred in our task. However, we believe the most important question in relation to our study is: to what extent did this learning influence our manipulations of interest?

In our reply to reviewer 1, we already showed that a re-analysis of the fMRI results using the trial-by-trial estimates of the omission contrasts revealed no Probability x Run interaction, suggesting that – overall – the probability effect remained stable over the course of the experiment. However, inspired by the alternative explanation that was proposed by this reviewer, we now also assessed the role of the Gambler’s fallacy in a separate set of analyses. Indeed, it is possible that participants start to expect a stimulation more after more time has passed since the last stimulation was experienced. To test this alternative hypothesis, we specified two new regressors that calculated for each trial of each participant how many trials had passed since the last stimulation (or since the beginning of the experiment) either overall (across all trials of all probability types; hence called the overall-lag regressor) or per probability level (across trials of each probability type separately; hence called the lag-per-probability regressor). For both regressors a value of 0 indicates that the previous trial was either a stimulation trial or the start of experiment, a value of 1 means that the last stimulation trial was 2 trials ago, etc.

The results of these additional analyses are added in a supplemental note (see supplemental note 6), and referred to in the main text (see lines 231-236: “Likewise, a post-hoc trial-by-trial analysis of the omission-related fMRI activations confirmed that the Probability effect for the VTA/SN activations was stable over the course of the experiment (no Probability x Run interaction) and remained present when accounting for the Gambler’s fallacy (i.e., the possibility that participants start to expect a stimulation more when more time has passed since the last stimulation was experienced) (see supplemental note 6). Overall, these post-hoc analyses further confirm the PE-profile of omission-related VTA/SN responses”.

Addition to supplemental material (pages 16-18)

Supplemental Note 6: The effect of Run and the Gambler’s Fallacy

A question that was raised by the reviewers was whether omission-related responses could be influenced by dynamical learning or the Gambler's Fallacy, which might have affected the effectiveness of the Probability manipulation.

Inspired by this question, we exploratorily assessed the role of the Gambler's Fallacy and the effects of Run in a separate set of analyses. Indeed, it is possible that participants start to expect a stimulation more when more time has passed since the last stimulation was experienced. To test this alternative hypothesis, we specified two new regressors that calculated for each trial of each participant how many trials had passed since the last stimulation (or since the beginning of the experiment) either overall (across all trials of all probability types; hence called the overall-lag regressor) or per probability level (across trials of each probability type separately; hence called the lag-per-probability regressor). For both regressors a value of 0 indicates that the previous trial was either a stimulation trial or the start of experiment, a value of 1 means that the last stimulation trial was 2 trials ago, etc.

The new models including these regressors for each omission response type (i.e., omission-related activations for each ROI, relief, and omission-SCR) were specified as follows:

(1) For the overall lag:

Omission response ~ Probability * Intensity * Run + US-unpleasantness + Overall-lag + (1|Subject).

(2) For the lag per probability level:

Omission response ~ Probability * Intensity * Run + US-unpleasantness + Lag-perprobability : Probability + (1|Subject).

Where US-unpleasantness scores were mean-centered across participants; "*" represents main effects and interactions, and ":" represents an interaction (without main effect). Note that we only included an interaction for the lag-per-probability model to estimate separate lag-parameters for each probability level.

The results of these analyses are presented in the tables below. Overall, we found that adding these lag-regressors to the model did not alter our main results. That is: for the VTA/SN, relief and omission-SCR, the main effects of Probability and Intensity remained. Interestingly, the overall-lag-effect itself was significant for VTA/SN activations and omission SCR, indicating that VTA/SN activations were larger when more time had passed since the last stimulation (beta = 0.19), whereas SCR were smaller when more time had passed (beta = -0.03). This pattern is reminiscent of the Perruchet effect, namely that the explicit expectancy of a US increases over a run of non-reinforced trials (in line with the gambler's fallacy effect) whereas the conditioned physiological response to the conditional stimulus declines (in line with an extinction effect, Perruchet, 1985; McAndrew, Jones, McLaren, & McLaren, 2012). Thus, the observed dissociation between the VTA/SN activations and omission SCR might similarly point to two distinctive processes where VTA/SN activations are more dependent on a consciously controlled process that is subjected to the gambler's fallacy, whereas the strength of the omission SCR responses is more dependent on an automatic associative process that is subjected to extinction. Importantly, however, even though the temporal distance to the last stimulation had these opposing effects on VTA/SN activations and omission SCRs, the main effects of the probability manipulation remained significant for both outcome variables. This means that the core results of our study still hold.

Next to the overall-lag effect, the lag-per-probability regressor was only significant for the vmPFC. A follow-up of the beta estimates of the lag-per-probability regressors for each probability level revealed that vmPFC activations increased with increasing temporal

distance from the stimulation, but only for the 50% trials (beta = 0.47, t = 2.75, p < .01), and not the 25% (beta = 0.25, t = 1.49, p = .14) or the 75% trials (beta = 0.28, t = 1.62, p = .10).

Author response table 1.

F-statistics and corresponding p-values from the overall lag model

(*) F-test and p-values were based on the model where outliers were rescored to 2SD from the mean. Note that when retaining the influential outliers for this model, the p-value of the probability effect was p = .06. For all other outcome variables, rescoring the outliers did not change the results. Significant effects are indicated in bold.

	Omission responses											
Regressor	Relief		SCR		VTA/SN (*)		Left vPut		NAC		vmPFC	
	F	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*
Probability	**30.04**	**<.001**	**4.90**	**<.01**	**3.59**	**<.05**	0.18	n.s.	0.88	n.s.	1.73	n.s.
Intensity	**620.62**	**<.001**	**106.65**	**<.001**	**7.81**	**<.001**	**3.88**	**<.05**	0.70	n.s.	**4.90**	**<.01**
Run	**9.71**	**<.001**	**15.56**	**<.001**	1.13	n.s.	0.76	n.s.	0.62	n.s.	0.44	n.s.
Probability x Intensity	**3.69**	**<.01**	1.54	n.s.	1.15	n.s.	1.39	n.s.	1.76	n.s.	0.70	n.s.
Probability x Run	1.13	n.s.	1.24	n.s.	1.02	n.s.	1.22	n.s.	1.74	n.s.	1.26	n.s.
Intensity x Run	1.94	.07	1.30	n.s.	1.94	.07	1.59	n.s.	1.01	n.s.	**2.41**	**<.05**
Probability x Intensity x Run	0.56	n.s.	0.87	n.s.	0.76	n.s.	0.71	n.s.	0.76	n.s.	0.83	n.s.
Overall-lag	2.56	n.s.	**4.68**	**<.05**	**11.30**	**<.001**	<0.01	n.s.	0.16	n.s.	<0.01	n.s.
US-unpleasantness	**29.60**	**<.001**	3.00	.096	1.84	n.s.	0.06	n.s.	3.44	.07	0.26	n.s.

Author response table 2.

Table 2 F-statistics and corresponding p-values from the lag per probability level model

(*) F-test and p-values were based on the model where outliers were rescored to 2SD from the mean. Note that when retaining the influential outliers for this model, the p-value of the Intensity x Run interaction was p = .05. For all other outcome variables, rescoring the outliers did not change the results. Significant effects are indicated in bold.

	Omission responses											
Regressor	Relief		SCR		VTA/SN (*)		Left vPut		NAC		vmPFC	
	F	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*	*F*	*P*
Probability	**23.07**	**<.001**	**4.44**	**<.05**	**3.28**	**<.05**	0.22	n.s.	0.18	n.s.	0.31	n.s.
Intensity	**625.52**	**<.001**	**107.49**	**<.001**	**7.33**	**<.001**	**3.88**	**<.05**	0.66	n.s.	**4.85**	**<.01**
Run	**8.63**	**<.001**	**13.91**	**<.001**	1.09	n.s.	0.72	n.s.	0.60	n.s.	1.09	n.s.
Probability x Intensity	**3.87**	**<.01**	1.33	n.s.	1.01	n.s.	1.40	n.s.	1.81	n.s.	0.63	n.s.
Probability x Run	1.63	n.s.	1.10	n.s.	1.25	n.s.	1.16	n.s.	1.60	n.s.	1.17	n.s.
Intensity x Run	2.08	.053	1.25	n.s.	1.84	.09	1.61	n.s.	1.02	n.s.	**2.55**	**<.05**
Probability x Intensity x Run	0.55	n.s.	0.90	n.s.	0.72	n.s.	0.71	n.s.	0.73	n.s.	0.84	n.s.
Lag per probability: Probability	1.51	n.s.	1.32	n.s.	**1.12**	**n.s.**	0.10	n.s.	0.96	n.s.	**4.14**	**<.01**
USunpleasantness	**29.69**	**<.001**	2.99	.097	1.63	n.s.	0.06	n.s.	3.34	.08	0.27	n.s.

> *As the authors mentioned in the rebuttal letter, "selecting participants only if their anticipatory SCR monotonically increased with each increase in instructed probability 0% < 25% < 50% < 75% < 100%, N = 11 participants", only ~1/3 of the subjects actually showed strong evidence for the validity of the instructions. This further raises the question of whether the instructed design, due to the interference of false instruction and the dynamic learning among trials, is solid enough to test the hypothesis .*

We agree with the reviewer that a monotonic increase in anticipatory SCR with increasing probability instructions would provide the strongest evidence that the manipulation worked. However, it is well known that SCR is a noisy measure, and so the chances to see this monotonic increase are rather small, even if the underlying threat anticipation increases monotonically. Furthermore, between-subject variation is substantial in physiological measures, and it is not uncommon to observe, e.g., differential fear conditioning in one measure, but not in another (Lonsdorf & Merz, 2017). It is therefore not so surprising that 'only' 1/3 of our participants showed the perfect pattern of monotonically increasing SCR with

increasing probability instructions. That being said, it is also important to note that not all participants were considered for these follow-up analyses because valid SCR data was not always available.

Specifically, N = 4 participants were identified as anticipation non-responders (i.e. participant with smaller average SCR to the clock on 100% than on 0% trials; pre-registered criterium) and were excluded from the SCR-related analyses, and N = 1 participant had missing data due to technical difficulties. This means that only 26 (and not 31) participants were considered for the post hoc analyses. Taking this information into account, this means that 21 out of 26 participants (approximately 80%) showed stronger anticipatory SCR following 75% instructions compared to 25% instructions and that 11 out of 26 participants (approximately 40%) even showed the monotonical increase in their anticipatory SCR (see supplemental figure 4). Furthermore, although anticipatory SCR gradually decreased over the course of the experiment, there was no Run x Probability interaction, indicating that the instructions remained stable throughout the task (see supplemental figure 3).

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *A more operational approach might be to break the trials into different sections along the timeline and examine how much the results might have been affected across time. I expect the manipulation checks would hold for the first one or two runs and the authors then would have good reasons to focus on the behavioral and imaging results for those runs.*

This recommendation resembles the recommendation by reviewer 1. In our reply to reviewer 1, we showed the results of a re-analysis of the fMRI data using the trial-by-trial estimates of the omission contrasts, which revealed no Probability x Run interaction, suggesting that – overall - the probability effect remained (more or less) stable over the course of the experiment. For a more in depth discussion of the results of this additional analysis, we refer to our answer to reviewer 1.

> ***Reviewer #3 (Public Review):***
>
> *Comments on revised version:*
>
> *The authors were extremely responsive to the comments and provided a comprehensive rebuttal letter with a lot of detail to address the comments. The authors clarified their methodology, and rationale for their task design, which required some more explanation (at least for me) to understand. Some of the design elements were not clear to me in the original paper.*
>
> *The initial framing for their study is still in the domain of learning. The paper starts off with a description of extinction as the prime example of when threat is omitted. This could lead a reader to think the paper would speak to the role of prediction errors in extinction learning processes. But this is not their goal, as they emphasize repeatedly in their rebuttal letter. The revision also now details how using a conditioning/extinction framework doesn't suit their experimental needs.*

We thank the reviewer for pointing out this potential cause of confusion. We have now rewritten the starting paragraph of the introduction to more closely focus on prediction errors, and only discuss fear extinction as a potential paradigm that has been used to study the role of threat omission PE for fear extinction learning (see lines 40-55). We hope that these adaptations are sufficient to prevent any false expectations. However, as we have mentioned in our previous response letter, not talking about fear extinction at all would also not make sense in our opinion, since most of the knowledge we have gained about threat omission prediction errors to date is based on studies that employed these paradigms.

Adaptation in the revised manuscript (lines 40-55):

“We experience pleasurable relief when an expected threat stays away1. This relief indicates that the outcome we experienced (“nothing”) was better than we expected it to be (“threat”). Such a mismatch between expectation and outcome is generally regarded as the trigger for new learning, and is typically formalized as the prediction error (PE) that determines how much there can be learned in any given situation2. Over the last two decades, the PE elicited by the absence of expected threat (threat omission PE) has received increasing scientific interest, because it is thought to play a central role in learning of safety. Impaired safety learning is one of the core features of clinical anxiety4. A better understanding of how the threat omission PE is processed in the brain may therefore be key to optimizing therapeutic efforts to boost safety learning. Yet, despite its theoretical and clinical importance, research on how the threat omission PE is computed in the brain is only emerging.

To date, the threat omission PE has mainly been studied using fear extinction paradigms that mimic safety learning 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). These (primarily non-human) studies have revealed that there are striking similarities between the PE elicited by unexpected threat omission and the PE elicited by unexpected reward.”

> *It is reasonable to develop a new task to answer their experimental questions. By no means is there a requirement to use a conditioning/extinction paradigm to address their questions. As they say, "it is not necessary to adopt a learning paradigm to study omission responses", which I agree with. But the authors seem to want to have it both ways: they frame their paper around how important prediction errors are to extinction processes, but then go out of their way to say how they can't test their hypotheses with a learning paradigm.*
>
> *Part of their argument that they needed to develop their own task "outside of a learning context" goes as follows:*
>
> *(1) "...conditioning paradigms generally only include one level of aversive outcome: the electrical stimulation is either delivered or omitted. As a result, the magnitude-related axiom cannot be tested."*
>
> *(2) "....in conditioning tasks people generally learn fast, rendering relatively few trials on which the prediction is violated. As a result, there is generally little intra-individual variability in the PE responses"*
>
> *(3) "...because of the relatively low signal to noise ratio in fMRI measures, fear extinction studies often pool across trials to compare omission-related activity between early and late extinction, which further reduces the necessary variability to properly evaluate the probability axiom"*
>
> *These points seem to hinge on how tasks are "generally" constructed. However, there are many adaptations to learning tasks:*
>
> *(1) There is no rule that conditioning can't include different levels of aversive outcomes following different cues. In fact, their own design uses multiple cues that signal different intensities and probabilities. Saying that conditioning "generally only include one level of aversive outcome" is not an explanation for why "these paradigms are not tailored" for their research purposes. There are also several conditioning studies that have used different cues to signal different outcome probabilities. This is not uncommon, and in fact is what they use in their study, only with an instruction rather than through learning through experience, per se.*

> *(2) Conditioning/extinction doesn't have to occur fast. Just because people "generally learn fast" doesn't mean this has to be the case. Experiments can be designed to make learning more challenging or take longer (e.g., partial reinforcement). And there can be intra-individual differences in conditioning and extinction, especially if some cues have a lower probability of predicting the US than others. Again, because most conditioning tasks are usually constructed in a fairly simplistic manner doesn't negate the utility of learning paradigms to address PEaxioms.*
>
> *(3) Many studies have tracked trial-by-trial BOLD signal in learning studies (e.g., using parametric modulation). Again, just because other studies "often pool across trials" is not an explanation for these paradigms being ill-suited to study prediction errors. Indeed, most computational models used in fMRI are predicated on analyzing data at the trial level.*

We thank the reviewer for these remarks. The “fear conditioning and extinction paradigms” that we were referring to in this paragraph were the ones that have been used to study threat omission PE responses in previous research (e.g., Raczka et al., 2011; Thiele et al. 2021; Lange et al. 2020; Esser et al., 2021; Papalini et al., 2021; Vervliet et al. 2017). These studies have mainly used differential/multiple-cue protocols where either one (or two) CS+ and one CS- are trained in an acquisition phase and extinguished in the next phase. Thus, in these paradigms: (1) only one level of aversive US is used; and (2) as safety learning develops over the course of extinction, there are relatively few omission trials during which “large” threat omission PEs can be observed (e.g. from the 24 CS+ trials that were used during extinction in Esser et al., the steepest decreases in expectancy – and thus the largest PE – were found in first 6 trials); and (3) there was never absolute certainty that the stimulation will no longer follow. Some of these studies have indeed estimated the threat omission PE during the extinction phase based on learning models, and have entered these estimates as parametric modulators to CS-offset regressors. This is very informative. However, the exact model that was used differed per study (e.g. Rescorla-Wagner in Raczka et al. and Thiele et al.; or a Rescorla- Wagner–Pearce-Hall hybrid model in Esser et al.). We wanted to analyze threat omission-responses without commitment to a particular learning model. Thus, in order to examine how threat omissionresponses vary as a function of probability-related expectations, a paradigm that has multiple probability levels is recommended (e.g. Rutledge et al., 2010; Ojala et al., 2022)

The reviewer rightfully pointed out that conditioning paradigms (more generally) can be tailored to fit our purposes as well. Still, when doing so, the same adaptations as we outlined above need to be considered: i.e. include different levels of US intensity; different levels of probability; and conditions with full certainty about the US (non)occurrence. In our attempt to keep the experimental design as simple and straightforward as possible, we decided to rely on instructions for this purpose, rather than to train 3 (US levels) x 5 (reinforcement levels) = 15 different CSs. It is certainly possible to train multiple CSs of varying reinforcement rates (e.g. Grings et al. 1971, Ojala et al., 2022). However, given that US-expectation on each trial would primarily depend on the individual learning processes of the participants, using a conditioning task would make it more difficult to maintain experimental control over the level of USexpectation elicited by each CS. As a result, this would likely require more extensive training, and thus prolong the study procedure considerably. Furthermore, even though previous studies have trained different CSs for different reinforcement rates, most of these studies have only used one level of US. Thus, in order to not complexify our task to much, we decided to rely on instructions rather than to train CSs for multiple US levels (in addition to multiple reinforcement rates).

We have tried to clarify our reasoning in the revised version of the manuscript (see introduction, lines 100-113):

"The previously discussed fear conditioning and extinction studies have been invaluable for clarifying the role of the threat omission PE within a learning context. However, these studies were not tailored to create the varying intensity and probability-related conditions that are required to systematically evaluate the threat omission PE in the light of the PE axioms. First, these only included one level of aversive outcome: the electrical stimulation was either delivered or omitted; but the intensity of the stimulation was never experimentally manipulated within the same task. As a result, the magnitude-related axiom could not be tested. Second, as safety learning progressively developed over the course of extinction learning, the most informative trials to evaluate the probability axiom (i.e. the trials with the largest PE) were restricted to the first few CS+ offsets of the extinction phase, and the exact number of these informative trials likely differed across participants as a result of individually varying learning rates. This limited the experimental control and necessary variability to systematically evaluate the probability axiom. Third, because CS-US contingencies changed over the course of the task (e.g. from acquisition to extinction), there was never complete certainty about whether the US would (not) follow. This precluded a direct comparison of fully predicted outcomes. Finally, within a learning context, it remains unclear whether brain responses to the threat omission are in fact responses to the violation of expectancy itself, or whether they are the result of subsequent expectancy updating."

> *Again, the authors are free to develop their own task design that they think is best suited to address their experimental questions. For instance, if they truly believe that omission-related responses should be studied independent of updating. The question I'm still left puzzling is why the paper is so strongly framed around extinction (the word appears several times in the main body of the paper), which is a learning process, and yet the authors go out of their way to say that they can only test their hypotheses outside of a learning paradigm.*

As we have mentioned before, the reason why we refer to extinction studies is because most evidence on threat omission PE to date comes from fear extinction paradigms.

> *The authors did address other areas of concern, to varying extents. Some of these issues were somewhat glossed over in the rebuttal letter by noting them as limitations. For example, the issue with comparing 100% stimulation to 0% stimulation, when the shock contaminates the fMRI signal. This was noted as a limitation that should be addressed in future studies, bypassing the critical point.*

It is unclear to us what the reviewer means with "bypassing the critical point". We argued in the manuscript that the contrast we initially specified and preregistered to study axiom 3 (fully predicted outcomes elicit equivalent activation) could not be used for this purpose, as it was confounded by the delivery of the stimulation. Because 100% trials aways included the stimulation and 0% trials never included stimulation, there was no way to disentangle activations related to full predictability from activations related to the stimulation as such.

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *I'm not sure the new paragraph explaining why they can't use a learning task to test their hypotheses is very convincing, as I noted in my review. Again, it is not a problem to develop a new task to address their questions. They can justify why they want to use their task without describing (incorrectly in my opinion) that other tasks "generally" are constructed in a way that doesn't suit their needs.*

For an overview of the changes we made in response to this recommendation, we refer to our reply to the public review.

We look forward to your reply and are happy to provide answers to any further questions or comments you may have.

https://doi.org/10.7554/eLife.91400.3.sa0