Reinstatement of cortical outcome representations during higher-order learning

Lennart Luettgau[1,2*], Emanuele Porcu[3], Claus Tempelmann[4], Gerhard Jocham[1,2]

[1]Biological Psychology of Decision Making, Institute of Experimental Psychology, Heinrich Heine University, Düsseldorf, Germany
[2]Center for Behavioral Brain Sciences, Otto-von-Guericke University, Magdeburg, Germany
[3]Department of Biological Psychology, Otto-von-Guericke University, Magdeburg, Germany
[4]Department of Neurology, Otto-von-Guericke University, Magdeburg, Germany

*Correspondence to: luettgau@hhu.de

Abstract

Higher-order learning mechanisms like second-order conditioning (SOC) may allow agents to infer value in environments with sparse rewards. Surprisingly little is known about the neural mechanism underlying value transfer in SOC. Here, we show that SOC may be supported by reinstatement of outcome representations in orbitofrontal cortex (OFC) and interactions of OFC with amygdala. Our data suggest a mechanism by which value is conferred to stimuli that were never paired with reinforcement.

Learning in naturalistic settings is characterized by infrequent direct encounters with reinforcement (Gewirtz and Davis, 2000). Instead, several stimuli or actions often need to be chained together, such that one stimulus serves as a proxy for another stimulus that might eventually predict reward (Gewirtz and Davis, 2000). Exploiting such statistical regularities would allow agents to infer the value of stimuli or actions that were never directly followed by reinforcement and support instrumental behavior in the absence of reinforcement. Second-order conditioning (SOC) is an example of higher-order learning that allows exactly such spread of value to stimuli that have never been directly paired with reinforcement (Sharpe et al., 2017). Despite its relevance for real-life learning phenomena (Gewirtz and Davis, 2000; Parkes and Westbrook, 2011) and decades of behavioral investigations (Barnet et al., 1991; Gewirtz and Davis, 2000; Parkes and Westbrook, 2011; Rizley and Rescorla, 1972), surprisingly little is known about the neural mechanism underlying the transfer of value in SOC (Parkes and Westbrook, 2011). Recent evidence suggests that acquisition of incentive properties by second-order stimuli (CS_2) is related to activity of mesencephalic dopamine neurons that encode a reward prediction error elicited by first-order conditioned stimuli (CS_1) (Maes et al., 2020). Presumably, SOC critically depends on convergence of neural patterns representing CS_2 and outcomes (unconditioned stimuli, US) in amygdala (Gewirtz and Davis, 2000, 1997; Parkes and Westbrook, 2011; Setlow et al., 2002) and hippocampus (Gilboa et al., 2014). However, it remains unclear, how exactly a “direct link” between CS_2 and US-related motivational value is formed at the neural level.

We hypothesized that, during SOC, presentation of the CS_1 would trigger reinstatement of the neural pattern representing the US with which it had been paired during first-order conditioning (FOC). This would allow linkage of “online” US and CS_2 representations by associative plasticity in amygdala (Gewirtz and Davis, 2000, 1997; Parkes and Westbrook, 2011; Setlow et al., 2002) and hippocampus (Gilboa et al., 2014) (see Supplementary Information for alternative proposals discussed in the literature).

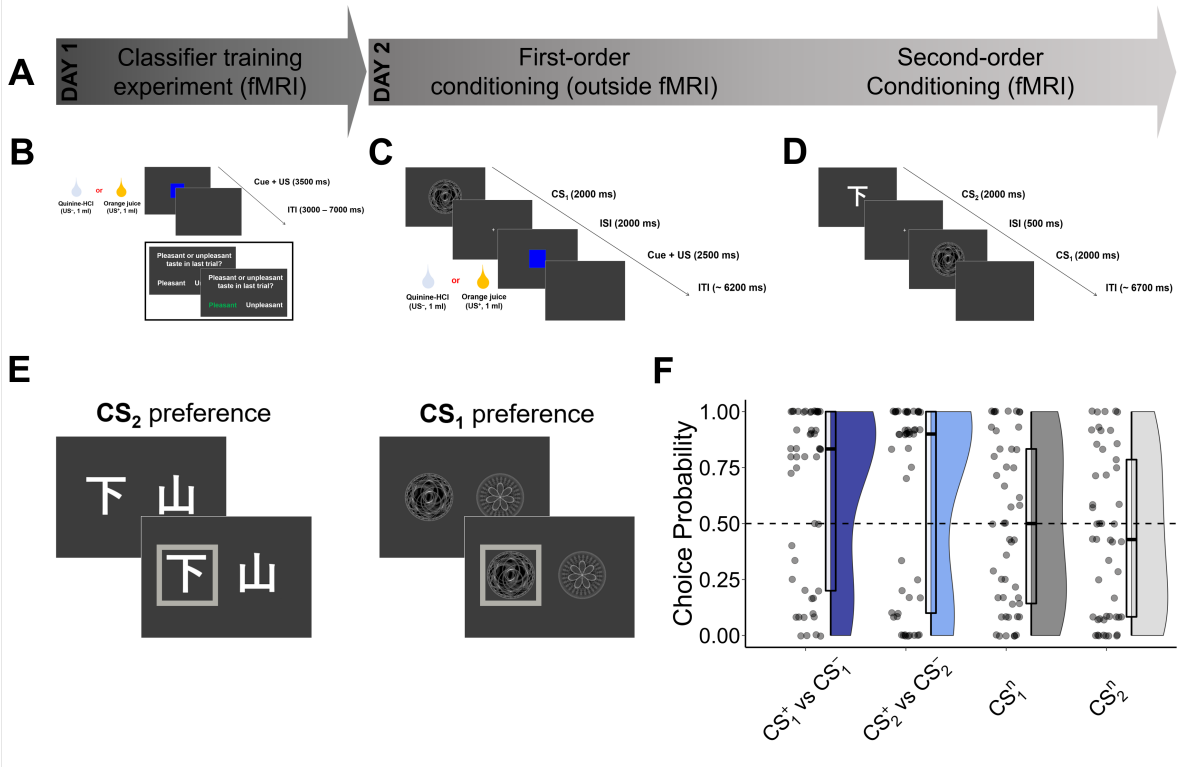

Fig. 1. Experimental procedure, task schematic and behavioral results.
A) Experimental procedure: Participants in the fMRI sample (*N* = 29) performed both day 1 and day 2 (B-E), participants in the behavioral sample (*N* = 20) only performed the procedures of day 2 (C-E, outside fMRI). B) Classifier training experiment: After rating of subjective value and intensity of both gustatory US (US^+: orange juice and US^-: quinine-HCL 0.1 mmol/l solved in purified water), in a total of five runs each US was administered twenty times (40 trials per run, 200 trials in total). Per trial, one US (1 ml bolus per trial) was delivered by a gustometer and infusion lines. The US bolus onset was preceded by a blue square. Trials were separated by an inter-trial-interval (ITI) marked by a grey screen. Participants performed a 0-back style attentional control task. After a pseudo-random 20% of trials, participants were presented with probe trials in which they were asked to indicate which US they had received last ("pleasant" (US^+) or "unpleasant" (US^-)). C) First-order conditioning: In each trial, a CS_1 was followed by an inter-stimulus interval marked by a fixation cross, and oral infusion of one US (1 ml bolus per trial). Trials were separated by an ITI marked by a grey screen. Each CS_1 was presented 50 times, amounting to 100 trials total. CS_1 were followed by a US with 80% probability. D) Second-order conditioning: In each trial, a CS_2 was followed by an inter-stimulus interval marked by a fixation cross, and a CS_1. CS_2 were followed by a CS_1 fully deterministically. Each trial was separated by an ITI marked by a grey screen. Each CS_2-CS_1 pair was presented 50 times (150 trials total). E) Choice preference test: Following SOC, participants were presented with two separate test phases consisting of repeated binary choices between pairs of CS_1 (right) and pairs of CS_2 (left) to assess behavioral signatures of first- and second-order conditioning. F) Behavioral results: Raincloud plots (Allen et al., 2019) showing density of choice probability data. CS_1^+ (dark blue, 89%-highest posterior density interval (89%-HPDI): [.58; .65]) and CS_2^+ (light blue, 89%-HPDI: [.60; .67]) were preferred over CS_1^- and CS_2^-, respectively. There was no evidence for choice probabilities of CS_1^n and CS_2^n being different from chance level in novel-to-lure comparison choice trials (dark and light gray, 89%-HPDIs: [.46; .52] and [.42; .48]). Box plot center lines represent sample medians and box bottom/top edges show $25^{th}/75^{th}$ percentile of the data, respectively.

To test these predictions, we used a second-order conditioning paradigm. Participants first established Pavlovian visuo-gustatory associations between a visual CS^+_1 and an appetitive gustatory US^+, and between a visual CS^-_1 and an aversive gustatory US^- during first-order

conditioning (FOC, Fig. 1C). During SOC, participants were visually presented with CS_2^+ followed by CS_1^+, CS_2^- followed by CS_1^-, and a novel pairing of CS_2^n followed by CS_1^n (Fig. 1D). Since CS_1^n had not been presented during first-order conditioning, both CS_2^n and CS_1^n were considered neutral CS, serving as control stimuli for mere exposure effects. Importantly, participants were not informed about the underlying associative structure of the tasks, which was aimed at leaving them unaware of the associative learning process. In a subsequent preference test (Fig. 1E), participants performed binary decisions between pairs of CS_1 and CS_2, interspersed with lure decisions. Participants showed a preference both for the appetitive first- and second order stimuli, CS_1^+ and CS_2^+, respectively, over the aversive CS_1^- and CS_2^- (Fig. 1F). Highest posterior density interval (HPDI) of choice probabilities (CP) for CS_1^+ versus CS_1^- [.58; .65] and for CS_2^+ versus CS_2^- [.60; .67] did not overlap with the defined region of practical equivalence (ROPE, CP = [.45; .55]), an interval of parameter values representing the null hypothesis of no difference from chance level. These findings show that our conditioning procedure conferred value to both CS_1 and CS_2. Importantly, there was no evidence for both choice probabilities of CS_1^n and CS_2^n being different from chance level in novel-to-lure comparison choice trials (Fig. 1F, HPDIs: [.46; .52] and [.42; .48], respectively). This rules out an explanation of CS_1^+ and CS_2^+ preference arising from mere-exposure or novelty.

We reasoned that a prerequisite for the observed associative transfer learning effect would be reinstatement of the neural US pattern by the paired CS_1 to establish a direct associative link between CS_2 and US. To test for cortical reinstatement of neural patterns representing US during second-order conditioning, we used functional magnetic resonance imaging (fMRI) and a cross-session (Stokes et al., 2009), cross-modality searchlight (Kriegeskorte et al., 2006) classification approach. To obtain unbiased estimates of the neural patterns representing our gustatory US, we first performed an fMRI classifier training experiment on day 1 (Fig. 1B) during which the US^+ and US^- were presented. A multivariate pattern classifier (linear support vector machine) was trained on the fMRI data from this session. On day 2, during SOC, we used the weights of these classifiers trained on gustatory

neural patterns to predict the identity of the visual CS (Fig. 1D) that had been paired with the US during FOC (Fig. 1C, see Methods and Fig. 2A for schematic of the classification approach). In other words, we used data from one sensory modality, assessed during a first day to train a classification model and tested the model's generalizability to unseen data from another sensory modality on a second day. We found evidence for reinstatement of US patterns in a region in the left lateral orbitofrontal cortex (lOFC). In this region, significant decoding of CS identity was possible on the basis of the US pattern obtained during classifier training on the previous day. Classification with a 3-mm searchlight revealed a small-volume corrected cluster of above-chance level (0.5) classification accuracy (Neurosynth meta-analytic functional mask, random-effect cluster-statistics (Stelzer et al., 2013)) in the left lateral orbitofrontal cortex (lOFC, Fig. 2B, z = 2.26, P = .012, corrected, one-tailed). Comparable results were obtained when using either an anatomical mask of lOFC or a functional mask from the literature (Supplementary Information). The location of this lOFC cluster is consistent with this region's well-documented role in gustatory processing, particularly in representing motivational (Rolls, 2006, 2000; Small et al., 1999) aspects of gustatory sensation and taste memory (Kobayashi et al., 2004). When we split participants into a high- and low-bias group, depending on their preference for the CS_2^+, we observed that significant classification of CS identity in lOFC was possible in the high bias group (Z = 3.43, P = .006, $U3_1$ = .88, one-sample Wilcoxon signed-rank test, two-tailed), but not in the low bias group (Z = 1.44, P = .151, $U3_1$ = .58, one-sample Wilcoxon signed-rank test, two-tailed, Fig. 2C).

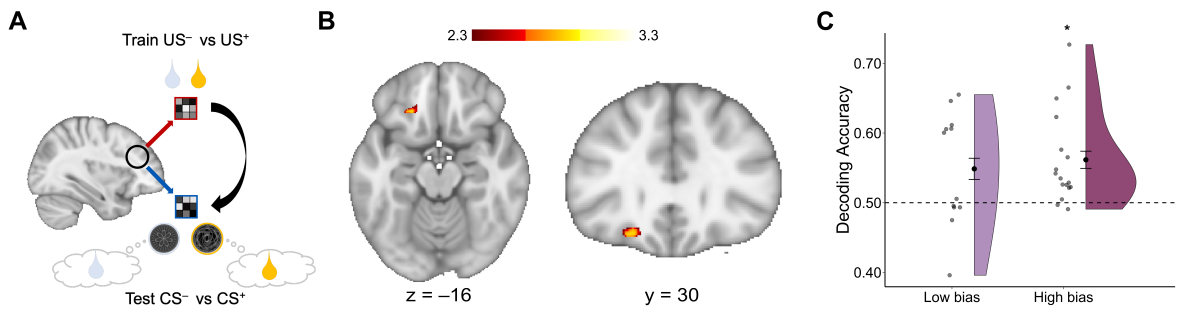

Figure 2. Cross-session, cross-modality classification analysis.
A) Analytic approach: For cross-session, cross-modality classification, a multivariate classifier was trained on the spatial activation patterns (*t* maps) of US^+ versus US^- and tested on CS_2^+/CS_1^+ versus CS_2^-/CS_1^- (CS_2^n/CS_1^n related activation were subtracted to control for general visual effects) during SOC within 3-mm searchlight spheres across the whole brain, resulting in one whole-brain map of classification accuracies per participant. Additionally, we repeated this procedure 100 times per participant with randomly permuted class labels in the training data set (US^+ versus US^-) to create chance level maps for random-effect cluster-statistics (Stelzer et al., 2013). B) Classification exceeded chance level (0.5) in a small-volume corrected cluster in the left lateral orbitofrontal cortex (lOFC, peak voxel at [x = –21, y = 30, z = –17], z = 2.26, P = .012, one-tailed, random-effect cluster-statistics, 50,000 iterations). C) Average classification accuracy in the lOFC cluster was significantly above chance level in the high bias group of participants showing higher than chance preference for CS_2^+ (dark purple raincloud, Z = 3.43, P = .006, $U3_1$ = .88, one-sample Wilcoxon signed-rank test, two-tailed), but not in the low bias group (light purple raincloud, Z = 1.44, P = .151, $U3_1$ = .58, one-sample Wilcoxon signed-rank test, two-tailed). Black dots indicate sample means and error bars represent standard errors of the means of the data, respectively.

To form an associative link between CS_2 and US, the reinstated US patterns need to be projected from their cortical storage site to regions like amygdala and hippocampus, allowing for convergence of US and CS_2 information. We investigated whole-brain functional covariation of the cluster in the left lOFC during SOC using a psychophysiological interaction (PPI) analysis. We contrasted CS^- and CS^+ trials, where the reinstated US is presumably paired with the CS_2, with CS^n trials. We found that functional covariation of BOLD signal in the left lOFC with a cluster in the hippocampus, extending to amygdala and medial temporal lobe (MTL) was higher in CS^- and CS^+ compared to CS^n trials (Fig. 3A, Z = 3.97, P = .029, whole-brain corrected). Furthermore, in the same PPI analysis, we found a positive correlation between second-order choice preference and functional covariation of the left lOFC with a region in the right lateral prefrontal cortex (Fig. 3B, Z = 4.21, P = .034, whole-brain corrected). The more participants preferred CS_2^+ (versus CS_2^-), the stronger these regions' BOLD signal covaried during SOC. If the amygdala uses the reinstated cortical outcome pattern in lOFC to acquire an association between CS_2 and the respective US, one

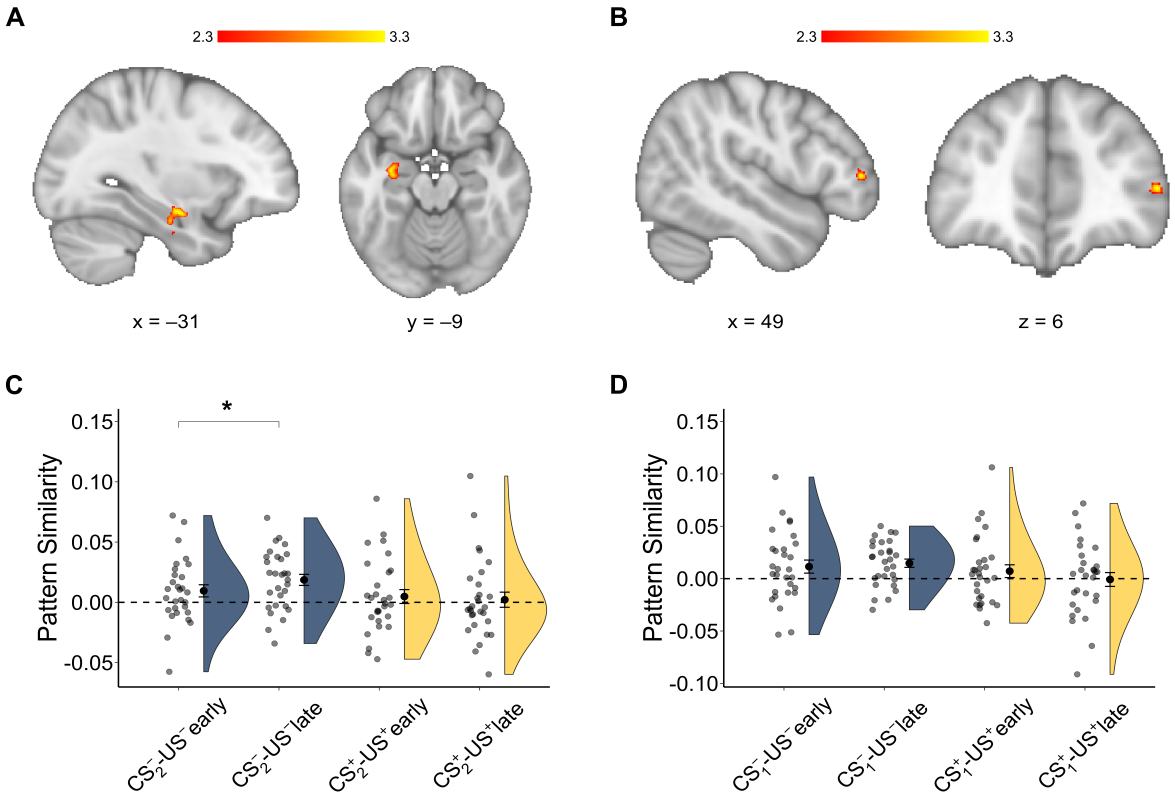

Figure 3. Psychophysiological interaction analysis and neural pattern similarity results. A) BOLD covariation in the left lOFC and a cluster in the anterior hippocampus, extending to (basolateral) amygdala and medial temporal lobe (MTL) was higher in CS^- and CS^+ trials than in CS^n trials (peak voxel at [x = –60, y = 3, z = –24], Z = 3.97, *P* = .029, whole-brain corrected). B) A positive correlation was observed between second-order choice preferences and functional covariation between the left lOFC and a region in the right lateral prefrontal cortex (peak voxel at [x = 49, y = 43, z = 6], Z = 4.21, *P* = .034, whole-brain corrected). The more participants preferred CS_2^+ (versus CS_2^-), the higher these regions' activation covaried during second-order conditioning. C & D) Template-based neural pattern similarity between classifier training experiment patterns of both US and least-squares separate (LS-S) estimated, pooled (early 25 and late 25 SOC trials) single-trial estimates of neural patterns representing CS_2 (C) and CS_1 (D), corrected for the respective pooled CSn neural patterns in a bilateral amygdala ROI. C) CS_2^- and US^- patterns became more similar from early to late trials (t_{28} = 1.88, *P* = .035, Cohen's *d* = .35, paired-samples t-test, one-tailed). However, there was also no evidence for a difference between early and late trial similarity for CS_2^+ and US^+ (t_{28} = .50, *P* = .310, Cohen's *d* = .09, paired-samples t-test, one-tailed). D) There was no evidence for change in similarity for CS_1^--US^- (t_{28} = .50, *P* = .310, Cohen's *d* = .09, paired-samples t-test, one-tailed) or CS_1^+ and US^+ neural pattern similarity (t_{28} = .50, *P* = .310, Cohen's *d* = .09, paired-samples t-test, one-tailed). Black dots indicate sample means and error bars represent standard errors of the means of the data, respectively.

would expect similarity between the neural patterns evoked by CS_2 and US, respectively. This similarity should increase over the course of second-order conditioning, as the association is being acquired. We first computed neural pattern similarity (Kriegeskorte et al., 2008) between the pattern evoked by CS_2 (during SOC) and their respective US (during classifier training) in the bilateral amygdala. There was significant pattern similarity between CS_2^- and US^- (t_{28} = 3.38, *P* = .002, $U3_1$ = .79; but not between CS_2^+ and US^+, *P* = .464, see

Supplementary Information). Notably, this similarity between CS_2^- and US^- increased from early to late trials of SOC (t_{28} = 1.88, P = .035, Cohen's d = .35, paired-samples t-test, one-tailed, Fig. 3C). This change in similarity was not present for the first-order CS_1^--US^- pattern similarity (t_{28} = .50, P = .310, paired-samples t-test, one-tailed, Fig. 3D). However, there was also no evidence for a difference between early and late trial similarity for CS_2^+ and US^+ or CS_1^+ and US^+ (all P > .635, paired-samples t-tests, one-tailed, Fig. 3C/D).

Our data provide evidence for transfer of outcome value during second-order conditioning, even when participants are unaware of the underlying associative structure, closely resembling findings in rodents (Maes et al., 2020; Sharpe et al., 2017). This shows that humans implicitly acquire preferences through higher-order transfer learning mechanisms, which goes beyond acquisition of explicit associative relationships between stimuli (Jara et al., 2006; Wang et al., 2020; Wimmer and Shohamy, 2012). Furthermore, we have shown reinstatement in lOFC of the neural outcome patterns evoked by visual first-order CS that had previously been paired with gustatory outcomes during first-order conditioning. This same lOFC region showed increased coupling with amygdala/hippocampus during second-order conditioning. Representations of visual second-order CS in amygdala/hippocampus became more similar to gustatory US representations from early to late phases of second-order conditioning, indicative of the gradual development of an associative link between CS_2^+ and US. Taken together, our data support the "direct link" hypothesis of second-order conditioning (Barnet et al., 1991; Gewirtz and Davis, 2000; Rizley and Rescorla, 1972), stating that second-order CS can be linked with the motivational value of outcomes by exploiting the relational structure of events. Our results suggest a neural mechanism by which the motivational value of outcomes can be propagated to stimuli that are never experienced in contiguity with reinforcement, allowing for credit assignment in real-world learning scenarios with infrequent direct encounters with rewards or punishments.

Methods

Participants

Participants were recruited from the local student community of the Otto-von-Guericke University Magdeburg, Germany by public advertisements and via online announcements. Only participants indicating no history of psychiatric or neurological disorder, no regular intake of medication known to interact with the central nervous system and participants not reporting quinine intolerance were included. Participants in both samples had normal or corrected-to-normal vision and did not report experience with Japanese kanjis or Chinese characters. All participants provided informed written consent before participation and received monetary compensation for taking part in the study. The study was approved by the local ethics committee at the medical faculty of the Otto von Guericke University Magdeburg, Germany (reference number: 101/15) and conducted in accordance with the Declaration of Helsinki.

32 healthy adult volunteers (age: M = 24.16, SD = 3.61 years, 15 males) participated in the fMRI study, 20 healthy adult volunteers (age: M = 23.40, SD = 3.07 years, 9 males), participated in the behavioral study. In the fMRI study, two participants were excluded from statistical analyses due to self-reports of having fallen asleep during the second-order conditioning scanning run. One additional participant had to be excluded due to a scanner malfunction and corruption of three of the five classifier training experiment scanning runs, thus leaving a total of n = 29 participants for final analyses.

Learning experiment – ratings

The learning experiment was performed on day 2 in the fMRI sample, in the behavioral sample, only the learning experiment was performed in a single session. Participants received written instructions for the experiment and were instructed once again on the computer screen. All experiments were programmed in MATLAB 2012b (MATLAB and Statistics Toolbox Release 2012b, The MathWorks, Inc., Natick, MA, USA), using Psychophysics Toolbox (Brainard, 1997) (version 3). Before and after the learning

experiment, participants rated ten different fractal images serving as first-order conditioned stimuli (CS_1), ten Japanese kanjis (Tamaoka et al., 2017) serving as second-order conditioned stimuli (CS_2) and gustatory stimuli (aversive: quinine-HCl 0.2 mmol/l solved in purified water, appetitive: either chocolate milk or orange juice (behavioral study, only orange juice in the fMRI study) serving as unconditioned stimuli (US). Ratings of subjective value/liking were assessed with the number buttons on a German (QWERTZ) computer keyboard from 1 (not liked) to 9 (very much liked). In the fMRI sample, participants additionally rated gustatory stimuli regarding their subjective intensity levels from 1 (low intense) to 9 (very intense). Three fractals and kanjis rated closest to 5 (equivalent to "neutral") were selected for first- and second order conditioning and their order was randomized before being associated with the US in first-order conditioning or with the first-order CS_1 in second-order conditioning. To ensure motivational salience of the gustatory US, participants were instructed to abstain from food for 12 hours and on average reported having fasted for 13.28 (*SD* = 2.71) hours. Participants reported intermediate levels of hunger before the task on a paper-pencil visual analog scale (VAS), ranging from 0 ("not hungry") to 100 ("very hungry"), *M* = 58.38 (*SD* = 29.61).

Learning experiment – first-order conditioning

In first-order conditioning (FOC), participants were presented with CS^+_1 followed by the appetitive US^+, and CS^-_1 followed by the aversive US^-. In a typical trial, a CS_1 (2000 ms) was followed by an inter-stimulus interval (1000 ms) marked by a fixation cross, and oral infusion of one US (1 ml bolus per trial). Each CS_1 was presented 50 times, amounting to 100 trials total. CS_1 were followed by a US with 80% probability (40 trials of each CS_1-US pair, 10 trials of CS_1-no US per CS). US were delivered by a MATLAB code-controlled custom-made gustometer consisting of two high pressure single syringe pumps (AL-1000HP, World Precision Instruments, Saratoga, FL) operating 50 ml Luer lock syringes. Syringes were attached to Luer lock infusion lines (1.40 m length, 2 mm inner diameter) that participants held centrally in their mouths like drinking straws. In the fMRI study, infusion line position

order (Q-O (*N* = 17) and O-Q (*N* = 15) for quinine (Q) and orange juice (O)) was counterbalanced across participants. In the fMRI study, the US bolus onset was preceded (500 ms) by a blue square that was presented for 2500 ms on the screen (Figure 1B). Participants were instructed to only swallow the US bolus upon offset of the blue square. Each trial was separated by an inter-trial-interval (ITI) marked by a grey screen. The ITI per trial was drawn from a discretized γ-distribution (shape = 6, scale = 1) truncated for an effective range of values between 3500 ms and 10,000 ms. Participants took self-paced breaks after each 10th trial during which they could drink water. Importantly, the instructions did not contain information about the underlying associative structure of the experiment, aiming at leaving participants unaware of the associative learning process. Instead, participants were instructed to perform a simple attentional control task, during which they should react as quickly and as correctly as possible by pressing the "y" button upon seeing the CS_1 colored in red. Each CS_1 was colored red in 10 % of the trials (90 % of trials grayscale image) and color did not predict US contingency. Performance during the task was rewarded with a bonus of 1 € (if > 70 % correct answers). Participants performed very well in the attentional control task (overall probability of correct answers: *M* = .99, *SD* = .05). In the fMRI sample, both ratings and first-order conditioning were performed outside the MRI scanner.

Learning experiment – second-order conditioning

For second-order conditioning (SOC), participants were presented with CS_2^+ followed by CS_1^+, CS_2^- followed by CS_1^-, and CS_2^n followed by CS_1^n. CS_1^n had not been presented during first-order conditioning and thus was not paired with any US. The novel CS_2^n-CS_1^n pair served to control for attentional and novelty effects. In each trial (Figure 1D), a CS_2 (2000 ms) was followed by an inter-stimulus interval (500 ms) marked by a fixation cross, and a CS_1 (2000 ms). CS_2 were followed by a CS_1 deterministically. Each trial was separated by an inter-trial-interval (ITI) marked by a grey screen. The ITI per trial was drawn from a discretized γ-distribution (shape = 7, scale = 1) truncated for an effective range of values

between 4000 ms and 10,000 ms. Each CS_2-CS_1 pair was presented 50 times, amounting to 150 trials total. Again, instructions did not explicitly mention relational structures to be learned, but participants were instructed to perform a simple attentional control task. Participants were instructed to respond as quickly and as correctly as possible by pressing the "y" button (behavioral sample) or with the right index finger on an MRI-compatible 4-button response box (fMRI sample) upon seeing the CS_2 tilted by a 45° angle. CS_2 were tilted in 10 % of the trials. Each CS_2 was tilted equally often. Performance during the task was rewarded with a bonus of 1 € (if > 70 % correct answers). Participants performed very well in the attentional control task (overall probability of correct answers: *M* = .97, *SD* = .07). In the fMRI study, SOC was performed inside the MRI scanner.

Learning experiment – choice preference test

Following SOC, participants were presented with two separate test phases consisting of repeated binary choices (30 trials) between pairs of CS_1 and pairs of CS_2 to assess behavioral signatures of first- and second-order conditioning. Each choice between CS_1^+ and CS_1^-, and between CS_2^+ and CS_2^- was presented ten times (twelve times in the fMRI sample). We hypothesized preference for both CS_1^+ and CS_2^+ over CS_1^- and CS_2^-, respectively, as expressed by choice probabilities above indifference criterion (CP > 0.5). Decision trials were interleaved with choices between the novel CS_1^n or CS_2^n and lure stimuli (fractal or kanjis, respectively) that had only been seen during pre-task rating in pseudo-random order. These novel-to-lure comparison trials were intended to rule out response biases related to mere exposure to the stimuli experienced during conditioning. We reasoned that if choices were reflecting acquired and transferred value and could not only be attributed to mere exposure-dependent response biases, choice probability in trials involving novel-to-lure stimuli should not exceed the indifference criterion. Choice options were presented for 1500 ms on the right- and left-hand side of the screen. Order (left/right) of choice options was counterbalanced. If participants did not respond within this time-window, a time-out message was displayed, and the respective trial was repeated at the end of the choice preference test.

Participants selected choice options by pressing the “y” (left option) or “m” (right option) button (behavioral sample, left or right index finger on MR-compatible response box in the fMRI sample). Participants were instructed to select the CS they preferred/liked more. Importantly, participants were never presented with the US related to their chosen or unchosen CS.

Behavioral analyses

Data were analyzed using MATLAB 2012b, 2015b and 2019a (The MathWorks, Inc., Natick, MA, USA) and RStudio (RStudioTeam, 2019, Boston, MA) using custom analysis scripts. We calculated an average choice probability (CP) per first- and second-order CS per subject including all binary decisions in which the respective CS was present (1 = selection of the respective target CS, 0 = selection of the alternative CS) for CS_1^+ and CS_2^+ (versus CS_1^- and CS_2^-, respectively). We hypothesized above-chance level CPs for both CS_1^+ and CS_2^+. For both first-order and second-order novel-to-lure comparison trials, we calculated an average CP per stimulus and expected no difference from chance level for these choices. For hypothesis testing, we calculated 89%-highest posterior density intervals (89%-HPDI) for each choice probability using the R package rethinking (McElreath, 2020). The HPDI represents the parameter values most consistent with the data. We specified a weakly informative prior probability distribution (10000 steps), an aggregated binomial distribution including the number of choices for the respective stimulus and the total number of choice trials involving the respective CS as likelihood function (Equation 1) and randomly drew 100000 samples from a grid-approximated (10000 steps) normalized posterior probability distribution. For hypothesis testing, we specified a region of practical equivalence (ROPE), i.e. an interval of parameter values (CP = [.45; .55]) representing the null hypothesis. Specifically, we expected that the 89%-HPDI would not overlap with the ROPE for choices of CS_1^+ and CS_2^+ (versus CS_1^- and CS_2^-, respectively), allowing to reject the null hypothesis. However, the ROPE would overlap with the 89%-HPDI for both first-order and second-order

novel-to-lure comparison trials. Overlap between HPDIs and ROPEs was quantified using the R package bayestestR (Makowski et al., 2019).

$$CP_i \sim \text{Binomial}(N_i, p_i)$$
$$p_i \sim \text{Normal}(0.5, 0.1) \quad (1)$$

where CP_i is the binomially distributed choice probability for CS_i (CS_1^+, CS_2^+, CS_1^n or CS_2^n). N_i is the total number of trials in which CS_i was present and p_i is the proportion of choices of CS_i.

fMRI – classifier training experiment

The classifier training experiment was always performed on day 1 of two consecutive testing days. This was done to acquire unbiased estimates of the neural patterns representing the two US for training of a multivariate classification algorithm (multivariate pattern analysis, MVPA), before any association of the US with first- or second order CS had been acquired. Participants received oral instructions for the classifier training experiment and were instructed once again on the projection screen in the MRI. Before and after the experiment, participants rated gustatory US (aversive: quinine-HCL 0.1 mmol/l solved in purified water and appetitive: orange juice) according to subjective value/liking and intensity via button presses with their right and left index and middle fingers on a MRI-compatible 4-button response box from left index (1 - not liked/not intense) to right middle finger (4 - very much liked/very intense). As on day 2, participants were instructed to abstain from food for 12 hours and on average reported to have fasted for 13.59 (*SD* = 2.18) hours. Participants reported intermediate levels of hunger before the experiment on a paper-pencil VAS, *M* = 48.50 (*SD* = 31.28). In each of the five total runs of the classifier training experiment, each US was administered twenty times (40 trials per run, 200 trials in total). Per trial, one US (1 ml bolus per trial) was delivered by a MATLAB code-controlled custom-made gustometer and Luer lock infusion lines (8.30 m length, 2 mm inner diameter). Participants held infusion lines centrally in their mouths like drinking straws. Additionally, infusion lines were fixed at absorbent materials attached to throat, chin and cheeks. The US bolus onset was preceded

(500 ms) by a blue square that was presented for 4000 ms on the screen. Participants were instructed to only swallow the US bolus upon offset of the blue square. Each trial was separated by an inter-trial-interval (ITI) marked by a grey screen. The intertrial interval was drawn from a uniform distribution with 5 discrete steps (range: 3000 – 7000 ms). Order of trials was pseudo-random, and possibly confounding low-level features of the schedule (e.g. number of same/different US repetitions, different ITI lengths following each US) were counterbalanced across US. During each run of the classifier training experiment, participants performed a 0-back style attentional control task. After a pseudo-random 20% of trials, participants were presented with probe trials in which they were asked to indicate which US they had received last ("pleasant" (US^+) or "unpleasant" (US^-)) via button presses with their right and left index fingers on an MRI-compatible response box. Correct responses were rewarded with 0.05 € and incorrect responses or time-out trials (without a response by the participant within 2500 ms after onset of the probe trial) resulted in a 0.05 € penalty which would be summed up as a bonus upon completion of the experiment. On average, participants earned a bonus of 1.86 € (*SD* = .11) during the classifier training experiment. Performance during the 0-back attentional control task was generally high (overall probability of correct answers, excluding time-out trials: *M* = .93, *SD* = .05).

fMRI – acquisition

During SOC, one run and during the classifier training experiment, five runs of fMRI were acquired on a 3 Tesla Siemens PRISMA MR-system (Siemens, Erlangen, Germany), using a 64-channel head coil. Blood oxygenation level dependent (BOLD) signals were acquired using a multi-band accelerated T2*-weighted echo-planar imaging (EPI) sequence (multi-band acceleration factor 2, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle = 80°, field of view (FoV) = 220 mm, voxel size = 2.2 × 2.2 × 2.2 mm, no gap). Per volume, 66 slices covering the whole brain, tilted by approximately 15° in z-direction relative to the anterior–posterior commissure plane were acquired in interleaved order. The first 5 volumes of the functional imaging time series were automatically discarded to allow for T1 saturation.

To ensure close spatial alignment of the acquired slices during both sessions, an online head scout sequence was measured during day 1 and the saved slice orientations were applied to day 2 measurements. At the end of both testing days, a B_0 magnitude and phase map was acquired to estimate field maps and B_0 field distortion during preprocessing (TR = 660 ms, TE 1 = 4.92 ms, TE 2 = 7.38 ms, flip angle = 60°, FoV = 220 mm). Additionally, before task-based fMRI on both days, a high-resolution three-dimensional T1-weighted anatomical map (TR = 2500 ms, TE = 2.82 ms, FoV = 256 mm, flip angle = 7°, voxel size = 1 × 1 × 1 mm, 192 slices) covering the whole brain was obtained using a magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence. This scan was used as anatomical reference to the EPI data during the registration procedure. For all cross-session classification analyses, SOC EPI data was referenced to orientation of the classifier training experiment.

fMRI – data preprocessing

All fMRI preprocessing steps were performed using tools from the Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL, v6.0 and v5.0) (Jenkinson et al., 2012). Preprocessing for each run of both classifier training task and SOC task included motion correction using rigid-body registration to the central volume of the functional time series (Jenkinson et al., 2002), correction for geometric distortions using the field maps and an n-dimensional phase-unwrapping algorithm (B_0 unwarping) (Jenkinson, 2003), slice timing correction using Hanning windowed sinc interpolation and high-pass filtering using a Gaussian-weighted lines filter of 1/100 Hz. EPI images were registered to the high-resolution structural image using affine linear registration (boundary-based registration) and then to standard (MNI) space using linear (12 degrees of freedom) and nonlinear registration (Andersson et al., 2007a, 2007b). Functional data was not spatially smoothed. We applied a conservative independent components analysis (ICA) to identify and remove obvious artefacts. Independent components were manually classified as signal or noise based on published guidelines (Griffanti et al., 2017), and noise components were removed from the

functional time series. General linear models (GLMs) were fitted into prewhitened data space to account for local autocorrelations (Woolrich et al., 2001).

fMRI – searchlight classification analyses

As conditioning enhances CS's ability to pre-activate neural ensembles representing an associated US, we expected that a multivariate classification algorithm trained on neural patterns representing the US (during the classifier training task) would be able to correctly predict the class of a paired CS during SOC. Hence, we used a cross-session (Stokes et al., 2009), cross-modality searchlight (Kriegeskorte et al., 2006) classification approach to identify brain regions in which the CS class during SOC could be predicted based on training the classifier during the classifier training experiment on day 1. For the searchlight classification analyses, the individual level (first level) GLM design matrix per run and participant of the classifier training experiment included four box-car regressors in total. Two regressors coded for onsets and durations of both US (each modelled as single events of 4000 ms duration) and two regressors coded onsets and durations of left and right button presses (delta stick functions on the recorded time of response button presses) and the six volume-to-volume motion parameters from motion correction during preprocessing were entered to account for residual head motion. The SOC run per participant included ten box-car regressors in total. Three regressors coding for onsets and durations of all combinations of CS_2-CS_1 (CS_2^+/CS_1^+, CS_2^-/CS_1^-, CS_2^n/CS_1^n) trials (each modelled as single events of 4500 ms duration, due to the 100 % contingency of CS-CS pairs), one regressor coding onsets and durations of right button presses (delta stick functions on the recorded time of response button presses), one regressor coding onset and duration of the within-run pause (45 sec), and the six volume-to-volume motion parameters from motion correction during preprocessing were entered.

Regressors were convolved with a hemodynamic response function (γ-function, mean lag = 6 s, SD = 3 s). Each first level GLM included two or three (for classifier training experiment or SOC, respectively) contrasts to estimate individual per run *t*-statistic maps for each US or each CS_2-CS_1 pair (for classifier training experiment or SOC, respectively). For SOC, the activation pattern of CS_2^n/CS_1^n was subtracted from CS_2^+/CS_1^+ and CS_2^-/CS_1^- activation patterns to account for general visual activation that was common to both CS and US (i.e., swallowing cue) presentation and could thus confound classification accuracies.

All classification-based analyses were conducted in subject native space. Per participant *t*-statistic maps were subjected to linear support vector machine (C-SVM, default C = 1) classification using the MATLAB-based multivariate pattern analysis toolbox CoSMoMVPA (Oosterhof et al., 2016). For cross-session, cross-modality classification, a classifier was trained on the spatial activation patterns of US^+ versus US^- and tested on CS_2^+/CS_1^+ versus CS_2^-/CS_1^- during SOC within 3-mm searchlight spheres across the whole brain. Each searchlight sphere classification accuracy was mapped to the center voxel of the sphere, resulting in one whole-brain map of classification accuracies per participant. Additionally, we repeated this procedure 100 times per participant with randomly permuted class labels in the training data set (US^+ versus US^-) to create chance level maps. Before group level statistics, the normalization parameters obtained from preprocessing were applied to both whole-brain classification accuracy maps and chance level maps for normalization to MNI space. The resulting normalized whole-brain maps were spatially smoothed with a Gaussian kernel (5 mm full-width half-maximum).

We employed small-volume correction (P_{SVC}) to assess significance of clusters with above-chance level (0.5) classification accuracy in the lateral orbitofrontal cortex (lOFC), our a priori region-of-interest (ROI). This region has been implicated in higher-order gustatory processing, such as motivational aspects and discrimination of taste (Kobayashi et al., 2004; Miranda, 2012; Small et al., 1999), but also representation and adaptive changes of stimulus-outcome associations (Boorman et al., 2016; Garvert et al., 2017; Jocham et al., 2016; Klein-Flugge et al., 2013; Luettgau et al., 2020). We thus reasoned that this brain region would be

a well-suited candidate for associative coupling between visual conditioned stimuli and gustatory outcomes and the proposed associative transfer learning. We employed two different approaches for ROI definition: 1) an independent anatomical mask of the lateral OFC (Harvard-Oxford Atlas) and 2) a meta-analysis-based explicit mask for functional activations related to the term “taste” in the Neurosynth data base (www.neurosynth.org), thresholded at $z > 7$. This value was chosen so that only meta-analytic clusters of activation, but not single voxels, survived thresholding. The Neurosynth-based mask encompassed, aside from the lateral OFC, functional activation clusters in bilateral anterior insula. This approach aimed at reducing inferential limitations related to arbitrary ROI definition.

Within these ROIs, we computed group-level random-effect cluster-statistics corrected for multiple comparisons in small volumes as implemented in CoSMoMVPA (Oosterhof et al., 2016). In brief, for 50,000 iterations, we randomly selected one chance level map per participant, calculated a group level z-statistic map for the respective permutation and finally compared the resulting cluster sizes drawn from this empirical null distribution (50,000 samples) to the clusters in the “real” classification accuracy map (Stelzer et al., 2013). Since classifications accuracies below chance level are generally limited in interpretability, we considered clusters significant if $P_{SVC} < .05$ (one-tailed).

Additionally, we split participants into a high- or low-bias group, depending on their preference for the CS_2^+ (vs. CS_2^-) and compared classification accuracies within the small-volume corrected cluster. The cluster ROI was built in MNI space and the ROI was then back-projected into subject native space using inverse normalization parameters obtained during preprocessing to extract individual averaged classification accuracies from the ROI. Both groups’ average extracted classification accuracies were separately tested against chance level (0.5) using one-sample Wilcoxon signed-rank tests. We report measures of effect size Cohen’s $U3_1$ for one-sample Wilcoxon signed-rank tests (range: 0 – 1, .5 indicating no effect), calculated in the MATLAB-based Measures-of-Effect-Size-toolbox (Hentschke, 2020).

fMRI – functional connectivity analyses

We investigated functional covariation between the MVPA-identified lOFC cluster surviving multiple comparisons (thresholded at z > 1.65, or *P* < .05, one-tailed) during SOC as a seed region and the whole-brain using a psychophysiological interaction (PPI) analysis in FSL. For each participant, we set up a first-level PPI GLM with the following regressors: 1) the BOLD timeseries (preprocessed functional time series) of the lOFC seed as physiological regressor, 2) onsets and durations of CS_2^+/CS_1^+ and CS_2^-/CS_1^- pairs (magnitude coded as 1) and CS_2^n/CS_1^n (magnitude coded as –1) as psychological regressor and 3) the interaction between 1) and 2) as psychophysiological interaction regressor. First-level regressor 2) was convolved with a hemodynamic response function (γ-function, mean lag = 6 s, SD = 3 s). Contrast images from the first level were then taken to the group level using a random effects analysis. In addition to the main effect of the PPI regressor, we also investigated whether task-related functional connectivity of lOFC with the rest of the brain was related to second-order conditioning. For this, we entered the averaged second-order choice preference test data per subject as an additional regressor to the group level GLMs. We used cluster-based correction with an activation threshold of Z > 2.3 and a cluster-extent threshold of *P* < .05 at whole-brain level to control for multiple comparisons.

fMRI – neural pattern similarity analyses

In addition to our classification-based approach, we investigated whether there was evidence for changes of neural pattern similarity across second-order conditioning. Specifically, we used a least-squares separate (LS-S) approach (Mumford et al., 2012) to deconvolve single-trial estimates of neural patterns representing CS_2 and CS_1 and performed a template-based neural pattern similarity analysis (Wimber et al., 2015) (a variant of representational similarity analysis, RSA (Kriegeskorte et al., 2008)) between patterns of the two US during the classifier training experiment and all trial-specific CS_2/CS_1 patterns during the SOC run. We estimated two subject-specific GLMs per trial, one for both CS_1 and CS_2, containing the following regressors: 1) onset and duration of the respective trial to be estimated (each

modelled as a single event of 2000 ms duration), 2) onsets and durations of the remaining CS_1 or CS_2 trials, respectively, 3) onsets and durations of all other CS_1 or CS_2 trials, respectively. Additionally, one regressor coding onsets and durations of right button presses (delta stick functions on the recorded time of response button presses), one regressor coding onset and duration of the within-run pause (45 sec), and the six volume-to-volume motion parameters from motion correction during preprocessing were entered. For example, the GLM of the 25th trial of CS_1^+, contained one regressor (onset and duration) modelling this particular trial and one separate regressor (onsets and durations) for all remaining CS_1 trials but the 25th trial of CS_1^+ (149) and one separate regressor (onsets and durations) modelling all CS_2 trials. This procedure is assumed to allow for better deconvolution of events happening within close temporal proximity than standard least-squares-based approaches, which generally do not distinguish between overall and trial-specific error terms (Abdulrahman and Henson, 2016; Mumford et al., 2012).

Each first level GLM included one contrast to model activation related to the respective trial of interest versus baseline. The a priori ROIs of the bilateral amygdala were built in MNI space and back-projected into subject native space using inverse normalization parameters obtained during preprocessing. We used these individual ROIs for spatially constrained multivoxel pattern extraction from the respective contrast *t*-value maps. Similarity-based analyses were carried out using CoSMoMVPA (Oosterhof et al., 2016). We employed 1−Pearson's product-moment correlation coefficient (1−*r*) as a measure of pairwise dissimilarity between trial-specific neural patterns and the US template. Within-subject pairwise neural dissimilarity was subtracted from 1 (to create a measure of neural pattern similarity) and Fisher-*Z* transformed to closer approximate normally distributed data. CS_1^n-US^-/CS_2^n-US^- and CS_1^n-US^+/CS_2^n-US^+ early and late neural pattern similarity were subtracted from CS_1^--US^-/CS_2^--US^- and CS_1^+-US^+/CS_2^+-US^+ early and late neural pattern similarity, respectively. We hypothesized that neural pattern similarity would show dissociable changes from early (first 25) to late (last 25) trials, as both CS_2^+ and CS_2^- should become more similar to the respective indirectly associated US, indicating associative

learning transfer from CS_1 to CS_2. Contrarily, CS_1^+ and CS_1^- should, if anything, become less similar to the respectively paired US (e.g. due to extinction). Average change of early to late CS-US neural pattern similarity was analyzed at the group level with paired-samples t-tests. Due to the expected negative change for CS_1^+ and CS_1^- and the expected positive change for CS_2^+ and CS_2^- across trials, we used one-tailed tests accordingly.

Acknowledgments

The authors thank all volunteers who participated in this study. The authors would like to thank Halla Mulla-Osman, Stefanie Linnhoff, Leonie Oevel and Denise Scheermann for their invaluable support during data acquisition. We thank the Center for Magnetic Resonance Research of the University of Minnesota for providing the Multiband accelerated fMRI sequence. Additionally, we thank Gunnar and Christian Lüttgau for electrical engineering support during construction of the gustometer, Karsta Benz for helpful discussions during the set-up of the experiment and Denise Kramer for production of the gustatory stimuli. **Funding:** This work was funded by the federal state of Saxony-Anhalt and the „European Regional Development Fund“ (ERDF 2014-2020), Vorhaben: Center for Behavioral Brain Sciences (CBBS), FKZ: ZS/2016/04/78113; **Competing interests:** Authors declare no competing interests

References

Abdulrahman, H., Henson, R.N., 2016. Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. Neuroimage 125, 756–766. doi:10.1016/j.neuroimage.2015.11.009

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: A multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. Wellcome Open Res. 4, 1–41. doi:10.12688/wellcomeopenres.15191.1

Andersson, J.L.., Jenkinson, M., Smith, S., 2007a. Non-linear registration aka spatial normalisation. FMRIB Tech. Rep. TRO7JA2 22.

Andersson, J.L.., Jenkinson, M., Smith, S.M., 2007b. Non-linear optimisation. FMRIB Tech. Rep. TRO7JA1.

Barnet, R.C., Arnold, H.M., Miller, R.R., 1991. Simultaneous conditioning demonstrated in second-order conditioning: Evidence for similar associative structure in forward and simultaneous conditioning. Learn. Motiv. 22, 253–268. doi:10.1016/0023-9690(91)90008-V

Boorman, E.D., Rajendran, V.G., O'Reilly, J.X., Behrens, T.E., 2016. Two Anatomically and Computationally Distinct Learning Signals Predict Changes to Stimulus-Outcome Associations in Hippocampus. Neuron 89, 1343–1354. doi:10.1016/j.neuron.2016.02.014

Brainard, D.H., 1997. The Psychophysics Toolbox. Spat. Vis. 10, 433–436.

Garvert, M.M., Dolan, R.J., Behrens, T.E., 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. Elife 6, 1–20. doi:10.7554/elife.17086

Gewirtz, J.C., Davis, M., 2000. Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. Learn. Mem. 7, 257–266. doi:10.1101/lm.35200

Gewirtz, J.C., Davis, M., 1997. Second-order fear conditioning prevented by blocking NMDA receptors in amygdala. Nature 388, 471–474. doi:10.1038/41325

Gilboa, A., Sekeres, M., Moscovitch, M., Winocur, G., 2014. Higher-order conditioning is impaired by hippocampal lesions. Curr. Biol. 24, 2202–2207. doi:10.1016/j.cub.2014.07.078

Griffanti, L., Douaud, G., Bijsterbosch, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M.F., Duff, E.P., Fitzgibbon, S., Westphal, R., Carone, D., Beckmann, C.F., Smith, S.M.,

2017. Hand classification of fMRI ICA noise components. Neuroimage 154, 188–205. doi:10.1016/J.NEUROIMAGE.2016.12.036

Hentschke, H., 2020. hhentschke/measures-of-effect-size-toolbox (https://www.github.com/hhentschke/measures-of-effect-size-toolbox), GitHub. Retrieved January 10, 2020.

Jara, E., Vila, J., Maldonado, A., 2006. Second-order conditioning of human causal learning. Learn. Motiv. 37, 230–246. doi:10.1016/j.lmot.2005.12.001

Jenkinson, M., 2003. Fast, automated,N-dimensional phase-unwrapping algorithm. Magn. Reson. Med. 49, 193–197. doi:10.1002/mrm.10354

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. Neuroimage 17, 825–841. doi:10.1006/NIMG.2002.1132

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. Neuroimage 62, 782–790. doi:10.1016/J.NEUROIMAGE.2011.09.015

Jocham, G., Brodersen, K.H.H., Constantinescu, A.O.O., Kahn, M.C.C., Ianni, A.M., Walton, M.E.E., Rushworth, M.F.F.S., Behrens, T.E.E.J., 2016. Reward-Guided Learning with and without Causal Attribution. Neuron 90, 177–190. doi:10.1016/j.neuron.2016.02.018

Klein-Flugge, M.C., Barron, H.C., Brodersen, K.H., Dolan, R.J., Behrens, T.E.J., 2013. Segregated Encoding of Reward-Identity and Stimulus-Reward Associations in Human Orbitofrontal Cortex. J. Neurosci. 33, 3202–3211. doi:10.1523/JNEUROSCI.2532-12.2013

Kobayashi, M., Takeda, M., Hattori, N., Fukunaga, M., Sasabe, T., Inoue, N., Nagai, Y., Sawada, T., Sadato, N., Watanabe, Y., 2004. Functional imaging of gustatory perception and imagery: “Top-down” processing of gustatory signals. Neuroimage 23, 1271–1282. doi:10.1016/j.neuroimage.2004.08.002

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868. doi:10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 1–28. doi:10.3389/neuro.06.004.2008

Luettgau, L., Tempelmann, C., Kaiser, L.F., Jocham, G., 2020. Decisions bias future choices by modifying hippocampal associative memories. bioRxiv.

doi:https://doi.org/10.1101/802462

Maes, E.J.P., Sharpe, M.J., Usypchuk, A.A., Lozzi, M., Chang, C.Y., Gardner, M.P.H., Schoenbaum, G., Iordanova, M.D., 2020. Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors. Nat. Neurosci. 23, 176–178. doi:10.1038/s41593-019-0574-1

Makowski, D., Ben-Schachar, M., Lüdecke, D., 2019. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. J. Open Source Softw. 4, 1541.

McElreath, R., 2020. rethinking: Statstical rethinking book package.

Miranda, M.I., 2012. Taste and odor recognition memory: The emotional flavor of life. Rev. Neurosci. 23, 481–499. doi:10.1515/revneuro-2012-0064

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59, 2636–2643. doi:10.1016/j.neuroimage.2011.08.076

Oosterhof, N.N., Connolly, A.C., Haxby, J. V, Rosa, M.J., 2016. CoSMoMVPA : Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab / GNU Octave 10, 1–27. doi:10.3389/fninf.2016.00027

Parkes, S.L., Westbrook, R.F., 2011. Role of the basolateral amygdala and NMDA receptors in higher-order conditioned fear. Rev. Neurosci. 22, 317–333. doi:10.1515/RNS.2011.025

Rizley, R.C., Rescorla, R.A., 1972. Associations in second-order conditioning and sensory preconditioning. J. Comp. Physiol. Psychol. 81, 1–11. doi:10.1037/h0033333

Rolls, E.T., 2006. Brain mechanisms underlying flavour and appetite. Philos. Trans. R. Soc. B Biol. Sci. 361, 1123–1136. doi:10.1098/rstb.2006.1852

Rolls, E.T., 2000. The Orbitofrontal Cortex and Reward. Cereb. Cortex 10, 284–294. doi:10.1093/cercor/10.3.284

RStudioTeam, 2019. RStudio: Integrated Development for R.

Setlow, B., Gallagher, M., Holland, P.C., 2002. The basolateral complex of the amygdala is necessary for acquisition but not expression of CS motivational value in appetitive Pavlovian second-order conditioning. Eur. J. Neurosci. 15, 1841–1853. doi:10.1046/j.1460-9568.2002.02010.x

Sharpe, M.J., Batchelor, H.M., Schoenbaum, G., 2017. Preconditioned cues have no value.

Elife 6. doi:10.7554/eLife.28362

Small, D.M., Zald, D.H., Jones-Gotman, M., Zatorre, R.J., Pardo, J. V., Frey, S., Petrides, M., 1999. Human cortical gustatory areas: A review of functional neuroimaging data. Neuroreport 10, 7–14. doi:10.1097/00001756-199901180-00002

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. Neuroimage 65, 69–82. doi:10.1016/J.NEUROIMAGE.2012.09.063

Stokes, M., Thompson, R., Cusack, R., Duncan, J., 2009. Top-down activation of shape-specific population codes in visual cortex during mental imagery. J. Neurosci. 29, 1565–1572. doi:10.1523/JNEUROSCI.4657-08.2009

Tamaoka, K., Makioka, S., Sanders, S., Verdonschot, R.G., 2017. www.kanjidatabase.com: a new interactive online database for psychological and linguistic research on Japanese kanji and their compound words. Psychol. Res. 81, 696–708. doi:10.1007/s00426-016-0764-3

Wang, F., Schoenbaum, G., Kahnt, T., 2020. Interactions between human orbitofrontal cortex and hippocampus support model-based inference. PLoS Biol. 18, 1–24. doi:10.1371/journal.pbio.3000578

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., Anderson, M.C., 2015. Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. Nat. Neurosci. 18, 582–589. doi:10.1038/nn.3973

Wimmer, G.E., Shohamy, D., 2012. Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. Science 338, 270–3. doi:10.1126/science.1223252

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. Neuroimage 14, 1370–1386. doi:10.1006/NIMG.2001.0931

Supplementary Information

Reinstatement of cortical outcome representations during higher-order learning

Luettgau et al.

Supplementary results

CS and US ratings

During the ratings prior to the learning experiment, the subjective values/liking of the three kanjis selected as CS_2 did not differ significantly from each other (main effect of stimulus: $F_{2,96} = 0.07$, $P = .993$, $\eta^2_p < .001$, rmANOVA). However, there was a significant main effect of stimulus for the three fractals selected as CS_1 (main effect of stimulus: $F_{2,96} = 3.25$, $P = .043$, $\eta^2_p = .06$, rmANOVA) that was driven by higher rating of CS_1^n than CS_1^- ($t_{48} = 2.26$, $P = .028$, paired-samples t-test). This difference most likely resulted from the selection process and the inherent rank-ordering of the CS according to subjective value ratings. However, there were no significant differences between CS_1^n and CS_1^+ ($t_{48} = 1.68$, $P = .101$, paired-samples t-test). Most importantly, no rating differences were observed between CS_1^- and CS_1^+ ($t_{48} = 0.44$, $P = .659$, paired-samples t-test). Overall, US valence ratings did not differ from pre to post rating across both US (main effect of time (pre/post): $F_{1,48} = 0.34$, $P = .564$, $\eta^2_p = .01$, rmANOVA), but the appetitive US^+ was – as expected – rated as significantly more pleasant than the aversive US^- (main effect of stimulus: $F_{1,48} = 466.64$, $P < .001$, $\eta^2_p = .91$, rmANOVA). There was also a significant interaction effect between stimulus and time, indicating that US^+ was rated as more pleasant and US^- was rated as less pleasant from pre to post (interaction effect of stimulus x time: $F_{1,48} = 7.32$ $P = .009$, $\eta^2_p = .13$, rmANOVA). Importantly, no subject rated US^- as more pleasant than US^+ (pre or post). In the fMRI study, intensity ratings did not differ between US overall (main effect of stimulus: $F_{1,28} = 1.05$, $P = .315$, $\eta^2_p = .04$, rmANOVA), but US were rated as significantly more intense at the post rating (main effect of time: $F_{1,28} = 4.89$, $P = .035$, $\eta^2_p = .15$, rmANOVA). There was also a significant interaction effect between stimulus and time, indicating that US^- intensity rating increased more strongly than US^+ from pre to post (interaction effect of stimulus x time: $F_{1,28} = 5.62$, $P = .025$, $\eta^2_p = .17$, rmANOVA). Average temperature (°Celsius) of US^+ ($M = 22.53$, $SD = 1.52$) and US^- ($M = 22.44$, $SD = 1.47$) did not differ before being loaded into the syringes ($t_{27} = 1.36$, $P = .184$, paired-samples t-test).

In the fMRI classifier training experiment, overall US valence ratings were higher in pre than in post rating (main effect of time (pre/post): $F_{1,28} = 11.67$, $P = .002$, $\eta^2_p = .29$, rmANOVA), but the appetitive US^+ was, as intended, rated as significantly more pleasant than the aversive US^- across ratings (main effect of stimulus: $F_{1,28} = 185.28$, $P < .001$, $\eta^2_p = .87$, rmANOVA). There was no interaction effect between stimulus and time ($F_{1,28} = 0.28$, $P = .602$, $\eta^2_p = .01$, rmANOVA). Importantly, no subject rated US^- as more pleasant than US^+ (pre or post). Intensity ratings for US^+ were higher than for US^- across ratings (main effect of stimulus: $F_{1,28} = 15.14$, $P < .001$, $\eta^2_p = .35$, rmANOVA). There was no difference between US ratings from pre to post (main effect of time: $F_{1,28} = 0.02$, $P = .899$, $\eta^2_p < .001$, rmANOVA).

However, there was also a marginal interaction effect between stimulus and time, indicating that US^- intensity rating increased while US^+ intensity rating decreased from pre to post (interaction effect of stimulus x time: $F_{1,28}$ = 3.93, P = .057, η^2_p = .12, rmANOVA). Average temperature of US^+ (M = 21.71, SD = 1.94°C) and US^- (M = 21.68, SD = 1.95) did not differ before being loaded into the syringes (t_{27} = 0.61, P = .544, paired-samples t-test).

fMRI searchlight classification

In the main article, we present evidence for reinstatement of US patterns in a region in the left lateral orbitofrontal cortex (lOFC). In this region, significant decoding of CS identity was possible on the basis of the US pattern obtained during classifier training on the previous day using classification with a 3-mm searchlight and small-volume correction in a ROI defined by a meta-analysis based on the term "taste" in the Neurosynth data base. To ensure that our results are not dependent on this particular choice of ROI, we repeated the same analysis using two different, independent ROI. The first was an anatomical mask of lateral orbitofrontal cortex (Harvard-Oxford atlas), the second was obtained from an independent gustatory mapping study by Benz et al. (personal correspondence, map as supplementary data). Using random-effect cluster-statistics (Stelzer et al., 2013) with 50,000 iterations, we found similar results in the left lOFC anatomical ROI (peak voxel at [x = –21, y = 30, z = –17], z = 1.96, P = .024, corrected, one-tailed) and in the mask from Benz et al. (peak voxel at [x = –21, y = 30, z = –17], z = 1.84, P = .033, corrected, one-tailed).

Supplementary Discussion

At least four behavioral and cognitive mechanisms accounting for second-order conditioning have been proposed (Barnet et al., 1991; Gewirtz and Davis, 2000; Parkes and Westbrook, 2011; Rizley and Rescorla, 1972). Two of these hypotheses state that CS_2 could become associated with a CR using the associative link between CS_2 and CS_1 or because the CS_1-evoked CR is directly paired with CS_2 during SOC. According to another hypothesis, CS_1 reactivates a representation of the associated US and the reinstated US evokes a CR that is subsequently paired with CS_2. Behavioral evidence suggests that neither of these three hypotheses can account for second-order learning effects, as for instance CS_1 extinction after second-order training does only extinguish CS_1-related CRs, but leaves expression of CRs toward CS_2 unaffected (Barnet et al., 1991; Rizley and Rescorla, 1972). Furthermore, the occurrence of CRs toward CS_2 despite absent CS_1-related CRs resulting from simultaneously pairing CS_1 and US in first-order conditioning renders a mediation of second-order conditioning by CS_1-CS_2 associations unlikely (Barnet et al., 1991). These findings lend only indirect support to the conclusion that CS_2 is paired directly with a neural representation of the US, or more precisely, with the motivational state or value conveyed by the US during second-order training ("direct link" hypothesis) (Gewirtz and Davis, 2000; Parkes and Westbrook, 2011).

We had first computed neural pattern similarity (Kriegeskorte et al., 2008) in the bilateral amygdala ROI between the pattern evoked by CS_2 (during second-order learning) and their respective US (during classifier training). Although there was significant pattern similarity between CS_2^- and US^- (t_{28} = 3.38, P = .002, $U3_1$ = .79, one-sample t-tests vs. 0, two-tailed), we did not find evidence for pattern similarity between CS_2^+ and US^+ (t_{28} = 0.74, P = .464, $U3_1$ = .55, one-sample t-tests vs. 0, two-tailed). We reason that this could be due to differential motivational aspects of the gustatory outcomes we used. The US^- consisted of an aversively tasting quinine solution, which has been shown to produce innate avoidance responses that are not thought to be generated by learning (Yiannakas and Rosenblum, 2017). In contrast, US^+ was orange juice, a compound stimulus that consists of many taste features (e.g. sour and fruity components), next to sweet sensation, which also produces innate responses (Yiannakas and Rosenblum, 2017). Presumably, this heterogeneity of taste facets in US^+ had induced more variability in taste responses in our participants. Additionally, it appears likely that in a society offering sugary and highly palatable foods and drinks in abundance, US^+ (an orange juice, most likely a very familiar taste to most of our participants) might not have been perceived rewarding and novel enough to represent a motivationally relevant and salient event for the participants in our experiment. It might well be that only the US^- (an aversively tasting quinine solution) was considered behaviorally and motivationally relevant by the participants in our experiment. One could speculate that participants in our

experiment did not have experience with the aversive taste of quinine, rendering it a novel and salient gustatory sensation. Thus, it is possible that the observed preference for CS_1^+ and CS_2^+ was indeed due to an aversion against CS_1^- and CS_2^-.

Supplementary References

Barnet, R.C., Arnold, H.M., Miller, R.R., 1991. Simultaneous conditioning demonstrated in second-order conditioning: Evidence for similar associative structure in forward and simultaneous conditioning. Learn. Motiv. 22, 253–268. doi:10.1016/0023-9690(91)90008-V

Gewirtz, J.C., Davis, M., 2000. Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. Learn. Mem. 7, 257–266. doi:10.1101/lm.35200

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 1–28. doi:10.3389/neuro.06.004.2008

Parkes, S.L., Westbrook, R.F., 2011. Role of the basolateral amygdala and NMDA receptors in higher-order conditioned fear. Rev. Neurosci. 22, 317–333. doi:10.1515/RNS.2011.025

Rizley, R.C., Rescorla, R.A., 1972. Associations in second-order conditioning and sensory preconditioning. J. Comp. Physiol. Psychol. 81, 1–11. doi:10.1037/h0033333

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. Neuroimage 65, 69–82. doi:10.1016/J.NEUROIMAGE.2012.09.063

Yiannakas, A., Rosenblum, K., 2017. The insula and taste learning. Front. Mol. Neurosci. 10, 1–24. doi:10.3389/fnmol.2017.00335