nature communications

Article

https://doi.org/10.1038/s41467-025-57872-w

Eating disorder symptoms and emotional arousal modulate food biases during reward learning in females

Received: 27 March 2024

Accepted: 4 March 2025

Published online: 26 March 2025

Check for updates

Nina Rouhani[1] ✉, Cooper D. Grossman[2], Jamie Feusner[3,4,5] & Anita Tusche[6,7]

Food seeking and avoidance engage primary reward systems to drive behavior. It is nevertheless unclear whether innate or learned food biases interact with general reward processing to interfere with goal-directed choice. To this end, we recruited a large non-clinical sample of females with high eating-disorder symptoms ('HED') and a matched sample of females with low eating-disorder symptoms ('LED') to complete a reward-learning task where the calorie content of food stimuli was incidental to the goal of maximizing monetary reward. We find and replicate a low-calorie food bias in HED and a high-calorie food bias in LED, reflecting the strength of pre-experimental food-reward associations. An emotional arousal manipulation shifts this group-dependent bias across individual differences, with interoceptive awareness predicting this change. Reinforcement-learning models further identify distinct cognitive components supporting these group-specific food biases. Our results highlight the influence of reinforcement-based mechanisms and emotional arousal in eliciting potentially maladaptive food-reward associations.

The visceral drive to eat is a primary reinforcer of the brain's reward system, motivating the seeking and consumption of food critical for survival[1]. Preference for energy-dense or high-calorie food is thought to indicate evolutionary adaptation, orienting cognitive processes, including attention, motivation, and memory mechanisms, towards food that is more likely to provide energy and satisfy hunger[2,3]. Nonetheless, food preferences and dietary behavior are linked to a range of clinically-relevant conditions, from obesity to eating disorders with high mortality rates, such as anorexia nervosa and bulimia nervosa[4]. The high prevalence of disordered eating[5], characterized by patterns of restricted eating or loss of control overeating, has triggered substantial medical attention, including a focus on psychotherapeutic, pharmaceutical and neuromodulatory interventions[6]. Research on the underlying cognitive mechanisms, which remains incomplete, may further help delineate the progression of aberrant eating behavior and offer cognitive targets for treatment.

Previous work has highlighted the role of reward and control mechanisms in supporting eating decisions, as a food-rich environment extends the function of feeding beyond homeostatic balance, risking the deployment of maladaptive hedonic pathways[7]. Overconsumption is linked to the neural signatures of substance use disorder and compulsive behavior[3,7,8]. Moreover, underconsumption or restricted eating may engage similar neural circuits to generate opposite behavior[9]. Of special interest are models of anorexia nervosa, as patients can develop the motivation to forgo food altogether, against all biological instincts. While control strategies may be required to initiate such behavior, repeatedly restricting eating (and experiencing its associated rewards) is thought to deploy dopaminergic circuits to form rigid habits[9,10]. Accordingly, food decisions in anorexia nervosa are supported by habit-specific regions in the brain[11–13], which likely reinforce the avoidance of high- versus low-calorie food. In fact, implicit preference for low- versus high-calorie

[1]Department of Psychology, University of Southern California, Los Angeles, CA 90089, USA. [2]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA. [3]Department of Psychiatry, University of Toronto, Toronto, ON M5T 1R8, Canada. [4]Centre for Addiction and Mental Health, Toronto, Toronto, ON, Canada. [5]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden. [6]Department of Psychology, Queen's University, Kingston, ON K7L 3N6, Canada. [7]Center for Neuroscience Studies, Queen's University, Kingston, ON K7L 3N6, Canada.
✉e-mail: nrouhani@usc.edu

food predicts disease severity in anorexia nervosa[14], and persists after weight restoration (regardless of explicit liking), with the opposite pattern in healthy controls[15].

The current study tested if and how innate or learned reward associations with high- or low-calorie food ('food bias') may interfere with goal-directed behavior in a task where the food stimuli were completely incidental to the goal of maximizing rewards. We additionally investigated how emotional arousal, thought to modulate decision-making by altering somatic states[16] as well as its subjective perception (interoception), may shift food biases. We examined two populations with putatively distinct food biases: those with active eating disorder symptoms and a self-reported history of anorexia and/or bulimia nervosa (i.e., high eating-disorder symptoms, 'HED' group; $n = 138$) and those without any symptoms or history of eating disorder (i.e., low eating-disorder symptoms, 'LED' group; $n = 141$). Computational models of choice data further identified which components of reinforcement learning were modulated by group-specific food biases and emotional arousal.

To examine food biases and their underlying cognitive mechanism, we designed a reward learning paradigm (Fig. 1A). Here, participants learned the values of two categories from trial and error with the objective of choosing the more (probabilistically) rewarding category to maximize monetary earnings. The categories were either food-related (food task: 'raw' versus 'cooked/baked' food, serving as a cover for low- versus high-calorie food, respectively) or unrelated to food (control task: indoor versus outdoor scenes). Participants completed both the food and control tasks, and we predicted that overall task performance would be similar between groups. We nevertheless expected a food-specific bias, such that the LED group would learn to associate monetary reward better with high- (versus low-) calorie food, while the HED group would show the opposite bias, associating reward more strongly with low- (versus high-) calorie food. Notably, we predicted these group-specific biases even though the calorie content of the food stimuli was irrelevant to the goal of choosing the more rewarding category.

Interestingly, emotional arousal and stress can also influence and even predict the trajectory of maladaptive eating habits[17]. Negative arousal, such as anxiety, as well as positive arousal, such as excitement, can both drive maladaptive eating patterns[7,18]. To examine the modulating role of affective states, our main study included an emotional arousal manipulation halfway through the reward learning task: participants won or lost \$8 by 'spinning' a wheel of fortune, either experiencing heightened positive emotion (win) or negative emotion (loss) after the spin[19] (50% likelihood of either outcome, Fig. 1A). We expected that heightened emotion, regardless of a positive or negative outcome, may enhance the salience of food stimuli and inhibit goal-relevant mechanisms, further enabling habit-related food biases in learning and decision-making.

Nonetheless, emotional arousal can be experienced differently given an individual's interoception, meaning their ability to perceive and interpret physiological signals from their body[20]. Deficits in interoception have been consistently reported across eating disorders, and are thought to drive maladaptive eating behaviors by suppressing hunger and satiety cues as well as general autonomic arousal systems[21,22]. Lower self-reported interoception has been especially pronounced in anorexia and bulimia nervosa (less so in binge-eating-disorder) in both current and recovered patients[23]. We therefore investigated whether group differences in self-reported arousal ratings and individual differences in interoceptive awareness predicted changes in food biases across the emotional arousal manipulation.

In this study, we find and replicate a food-related bias during reward learning: participants with high eating disorder symptoms better associate low- (versus high-) calorie food with monetary reward, while participants with low eating disorder symptoms better associate high- (versus low-) calorie food with monetary reward. Emotional arousal moreover homogenizes the group-dependent food bias across individual differences in interoceptive awareness. Reinforcement learning models highlight pre-experimental (innate or learned) food values and food-specific learning rates for positive (prediction-error) outcomes in supporting the observed food biases.

Results

Similar overall learning between groups and tasks

As our measure of learning and task performance, we considered 'correct' responses as the choice for the more rewarding category (regardless of whether the outcome on that trial was probabilistically rewarded or not). As expected, the LED and HED groups did not significantly differ in their learning performance (mixed-effects logistic regression predicting correct choice as a function of group across all tasks and blocks: B = 0.06, z = 0.84, $p = 0.400$; M = 0.06 [−0.08, 0.19]), nor in their ability to learn in the food versus the control task (interaction between group and task: B = −0.04, z = −0.68, $p = 0.495$; M = −0.04 [−0.16, 0.08]; Fig. 2A). We furthermore did not find significant group differences in either the food or control task analyzed separately (food task: B = 0.03, z = 0.43, $p = 0.668$; M = 0.03 [−0.12, 0.19]; Fig. 2C; control task: B = 0.08, z = 1.04, $p = 0.298$; M = 0.08 [−0.06, 0.22]; Fig. 2D). There were therefore no statistically significant group differences in the overall ability to perform the task.

Reward learning was guided by group-specific food biases

While both groups performed equally well on average, we identified group-specific biases in the expected direction in the food task: participants in the HED group performed better when the low- (versus high-) calorie food category was more rewarding, and vice versa for the LED group (across all food blocks, mixed-effects logistic regression predicting correct choice as a function of the interaction between group and reward category, B = −0.23, z = −2.19, $p = 0.0288$; M = −0.23 [−0.44, −0.03]). This pattern of results replicated our pre-registered pilot results, which did not include an emotional arousal manipulation in the reward learning task (for visual comparison to the pilot study, see SI 5 and Fig. S10A-B). Importantly, this bias was food specific: We did not find a significant group-specific category bias (for indoor or outdoor scenes) in the control task (interaction between group and reward category in control task, B = −0.11, z = −1.08, $p = 0.279$; M = −0.10 [−0.30, 0.10]), which was significantly different from the food task (three-way interaction between group, reward category and task, B = −0.33, z = −3.67, $p < 0.001$; M = −0.33 [−0.50, −0.15]). These findings demonstrate that food-reward associations can selectively influence general reward learning processes, even if the rewards are unrelated to actual food consumption.

Heightened emotional state modulated group-specific food biases

The emotional arousal manipulation increased alertness across groups (mixed-effects linear regression predicting alertness ratings as a function of (pre- versus post-) arousal block, B = 0.26, t = 4.53, $p < 0.001$; M = 0.26 [0.15, 0.37]; Fig. 3B; for visualization and analysis of all ratings, see SI 3.2 and Fig. S5). We did not find significant differences in arousal ratings between groups (ratings as a function of group, B = 0.08, t = 0.67, $p = 0.507$; M = 0.07 [−0.15, 0.30]), indicating that we did not replicate lower arousal ratings in the HED group from the pilot study (SI 5.3 and Fig. S10C-D). We further did not find the emotional arousal manipulation to significantly influence overall learning performance in either the food task (mixed-effects logistic regression predicting correct choice as a function of arousal block, B = 0.02, z = 0.50, $p = 0.616$; M = 0.03 [−0.08, 0.12]) or in the control task (B = 0.03, z = 0.69, $p = 0.491$; M = 0.03 [−0.06, 0.12]; Fig. 2B).

We instead found the emotional arousal manipulation to generate the group-specific food bias across participants in the food task (three-

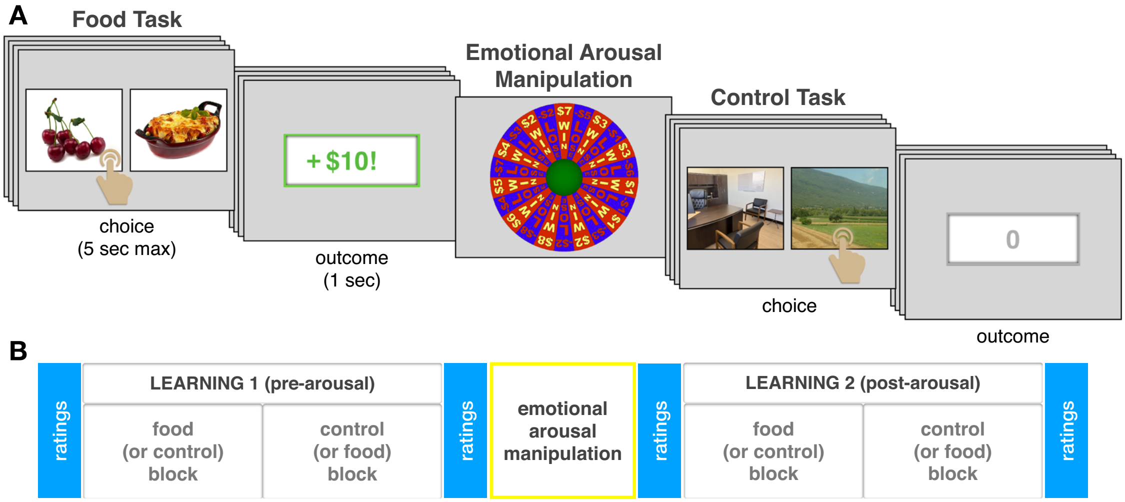

Fig. 1 | Experimental study design. A An example of a sequence of trials and (**B**) the overall structure of the reward learning task in the main study. Participants learned through trial and error which of two categories was more rewarding in each of 4 blocks (outcomes were $10 or $0); reward contingencies were probabilistic (0.75/0.25 reward probability); half the blocks included food stimuli ('food task', categories: 'raw' and 'cooked/baked' food, which served as a cover for low- and high-calorie food, respectively) or scene stimuli ('control task', categories: indoor and outdoor scenes). After completing one block of each task, participants experienced an emotional arousal manipulation in which they 'spun' a computerized wheel of fortune (0.50 probability of winning or losing $8). They next completed two more learning blocks of each task (order of blocks counterbalanced across participants). Participants completed self-ratings throughout the reward learning task (indicating level of 'alertness' (arousal), as well as hunger, thirst, and happiness) to check for changes induced by the emotional arousal manipulation. Food images (catalog numbers 280 and 143; Blechert J, Lender A, Polk S, Busch NA and Ohla K (2019). Food-Pics_Extended—An Image Database for Experimental Research on Eating and Appetite: Additional Images, Normative Ratings and an Updated Review. Front. Psychol. 10:307. https://doi.org/10.3389/fpsyg.2019.00307[64]) are released under a Creative Commons-Attribution-NonCommercial-Sharealike 3.0 license (https://creativecommons.org/licenses/by-nc-sa/3.0/).

way interaction between group, reward category, and arousal block: B = −0.45, z = −3.46, $p < 0.001$; M = −0.45 [−0.69, 0.19]; Fig. 3A versus 3 C). This effect was specific to the food (versus control) task (four-way interaction between group, reward category, arousal block, and task: B = −0.86, z = −4.87, $p < 0.001$; M = −0.87 [−1.24, −0.51]). When testing the food blocks separately, we did not find a significant group-specific food bias in the pre-arousal block (interaction between group and reward category, B = −0.02, z = −0.10, $p = 0.921$; M = −0.01 [−0.38, 0.35]), although there were marked individual differences (e.g., in interoceptive awareness, see below). The overall group-specific food bias emerged in the post-arousal block (interaction between group and reward category, B = −0.47, z = −2.55, $p = 0.0109$; M = −0.48 [−0.84, −0.11]).

We next examined whether the valence of the wheel-of-fortune outcome influenced learning. Consistent with previous work[19,24], we found a win on the wheel of fortune to increase happiness ratings (mixed-effects linear regression predicting happiness ratings as a function of (pre- versus post-) arousal block, B = 0.91, t = 12.20, $p < 0.001$; M = 0.91 [0.76, 1.06]; Fig. S5D), while a loss decreased happiness ratings (B = −1.21, t = −14.73, p < 0.001; M = −1.22 [−1.38, −1.05]; Fig. S5F; loss (versus win) outcome led to a stronger change in ratings, see SI 3.2), with no statistically significant difference in these changes between groups (interaction between group and arousal block, B = 0.08, t = 0.53, $p = 0.598$; M = 0.08 [−0.23, 0.39]; Fig. S5B). Interestingly, we did not find the valence of the wheel of fortune's outcome ($8 win or $8 loss) to have a statistically significant effect on the group-specific food bias (mixed-effects logistic regression predicting correct choice as a function of a four-way interaction between group, reward category, arousal block, and wheel-of-fortune outcome, B = 0.88, z = 1.53, $p = 0.127$; M = 0.93 [−0.22, 2.13]). These findings point to a general effect of heightened emotion (rather than a statistically significant effect of positive versus negative emotion), in shifting food biases. Moreover, we did not find the valence of the outcome to significantly influence overall learning performance in either the control or food tasks (for visualization and analysis, see SI 3.3 and Fig. S6).

Interoceptive awareness predicted food biases

Given deficits in interoceptive awareness in eating disorders[23], we tested whether individual differences in overall interoceptive awareness predicted food bias. We indeed found that interoceptive awareness (Multidimensional Assessment of Interoceptive Awareness, 'MAIA'[25]) modulated the change in the group-specific food bias across the emotional arousal manipulation (mixed-effects logistic regression predicting correct choice as a function of a four-way interaction between group, reward category, arousal block, and interoceptive awareness score, B = −0.61, z = −4.54, $p < 0.001$; M = −0.61[−0.89, −0.35]). Interoceptive awareness did not predict such changes in the control task (four-way interaction between group, reward category, arousal block, and interoceptive awareness score, B = 0.19, z = 1.40, $p = 0.163$; M = 0.18 [−0.08, 0.44]), which was significantly different from the food task (five-way interaction between group, reward category, arousal block, interoceptive awareness score, and task: B = −0.78, z = −4.21, $p < 0.001$; M = −0.78 [−1.15, −0.44]).

For visualization, we split our participant sample by the median interoceptive awareness score across LED and HED groups (median = 20.47; Fig. 3D–I). Higher interoceptive awareness additionally predicted higher self-reported alertness (a proxy for arousal) overall (mixed-effects linear regression predicting ratings as a function of interoceptive awareness score, B = 0.15, t = 3.09, $p = 0.00221$; M = 0.16 [0.05, 0.25]), with no statistical significant difference between HED and LED groups (interaction between group and interoceptive awareness score, B = 0.16, t = 1.57, $p = 0.117$; M = 0.16 [-0.04, 0.37]; Fig. 3E, H). To note, while interoceptive awareness was, on average, lower in the HED group (median = 19.72) versus the LED group (median = 21.78; B = 2.36, t = 3.13, $p = 0.00194$; M = 2.36 [0.93, 3.81]), there was substantial overlap between the groups' distributions of scores, see SI Fig. S1A.

Next, we examined the influence of individual interoception on the pre- and post-arousal blocks of the food task separately. We found interoceptive awareness to predict differences in food biases before the emotional arousal manipulation (three-way interaction between group, reward category, and interoceptive awareness score, B = 0.54,

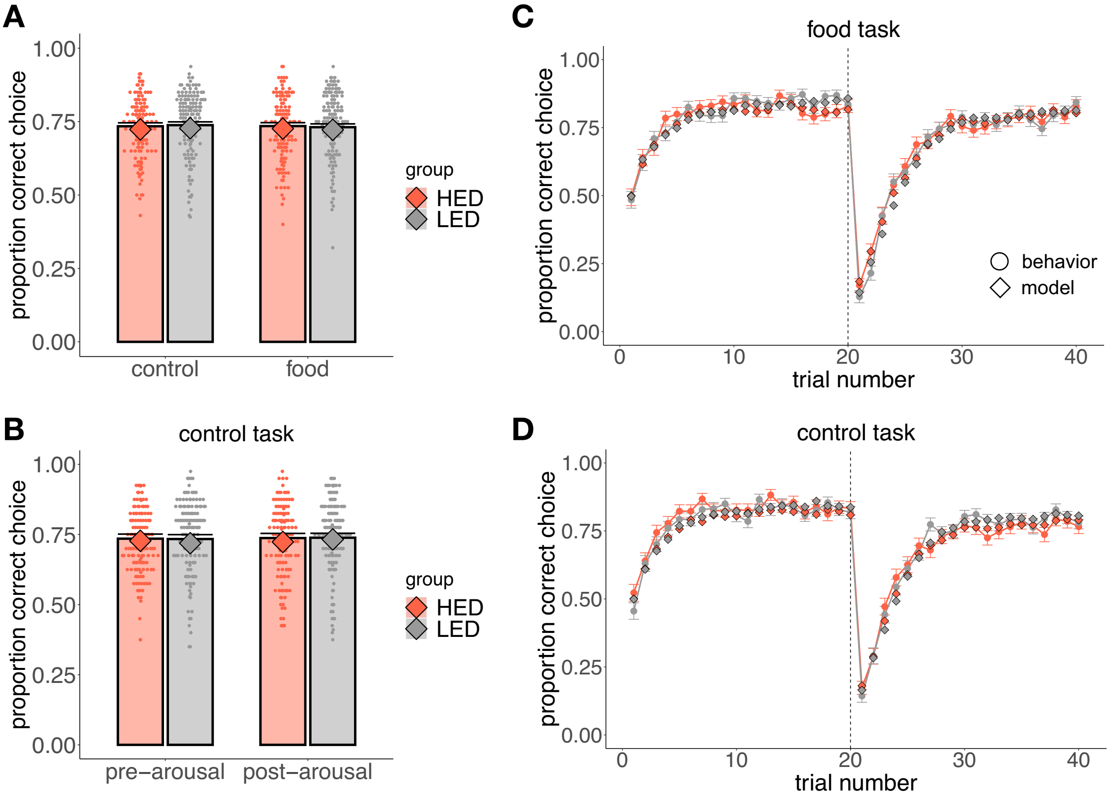

Fig. 2 | Participant- and model-generated behavior during reward learning in the high eating-disorder ('HED', red) and low eating-disorder ('LED', gray) symptom groups. A The average proportion of correct choice for the more rewarding category in the control and food tasks (across blocks), and **B** within the control task between the first block (before the emotional arousal manipulation, 'pre-arousal') and second block (after the emotional arousal manipulation, 'post-arousal'); participants are represented by dots. **C** The average trial-by-trial proportions of correct choice within the food task and **D** control task (dotted line at trial 20 indicates a 'change-point' event where there was a switch in the rewards associated with each category). The LED and HED groups did not differ in overall learning performance within or across (food/control) tasks and (pre-/post-arousal) blocks. Diamond points indicate average model predictions from the best-fitting model across participants. Error bars represent the standard error of the mean.

$z = 2.75$, $p = 0.00593$; M = 0.54 [0.14, 0.93]), but not after it (B = -0.16, z = -0.84, $p = 0.401$; M = -0.16 [−0.56, 0.24]). Moreover, when grouping participants by lower versus higher interoceptive awareness (Fig. 3D–I), overall HED and LED food biases were immediately apparent in participants with lower interoceptive awareness (interaction between (HED versus LED) group and reward category, B = −0.62, z = −2.30, $p = 0.0217$; M = −0.63 [−1.17, −0.07]; Fig. 3D), but not in those with higher interoceptive awareness, who demonstrated flipped food biases (B = 0.59, z = 2.22, $p = 0.0264$; M = 0.59 [0.05, 1.12]; Fig. 3G). In fact, higher interoceptive awareness predicted a food bias for high-calorie (instead of low-calorie) food in the HED group and vice-versa for the LED group.

Participants with higher interoceptive awareness further experienced a shift in their food biases after the emotional arousal manipulation (three-way interaction between group, reward category, and arousal block, B = −1.17, z = −6.26, $p < 0.001$; M = −1.17 [−1.54, −0.79]), whereas those with lower interoceptive awareness did not demonstrate a significant shift (B = 0.22, z = 1.14, $p = 0.254$; M = 0.22 [−0.16, 0.60]). We therefore found that individuals with higher interoception demonstrated an initial reversal of potentially harmful (to HED or LED) food-reward associations, but they were also the most affected by the emotional arousal manipulation. These results point to a potential intervention (increasing interoceptive awareness) as well as a challenge (emotional arousal) to reversing maladaptive food-reward associations.

In an exploratory analysis (SI 1.5), we additionally determined which of eight MAIA subscales (SI Fig. S4) supported this effect of interoceptive awareness (correcting for multiple comparisons). We found that four subscales modulated the group-specific food bias across the emotional arousal manipulation (all in the same direction as the overall interoceptive awareness score). The predictive subscales were related to actively engaging with and trusting physical and emotional signals from the body (attention regulation, body listening, trust, and emotional awareness), whereas the remaining four subscales – related to simply noticing body signals (noticing) or regulating negative emotional states (not-distracting, not-worrying, self-regulation)—were not predictive (for full analysis, see SI 1.5).

Group-specific food biases were captured in reinforcement learning models

To identify the precise cognitive mechanism underlying food biases in reward learning, we fit hierarchical reinforcement learning models to choice data (see Methods for model specifications). We used standard Q-learning[26] as our baseline model. Within this reinforcement learning framework, we examined four cognitive components that could support group biases in decision-making: (1) sensitivity to reward receipt, meaning the extent to which the reward outcome is valued[27], (2) distinct learning rates for positive versus negative prediction errors ('positive/negative learning rates'), referring to the extent to which positive versus negative outcomes asymmetrically update values[28,29], (3) persistence in choosing the same category that was previously chosen[30], and (4) initial values for each reward category, reflecting pre-experimental and potentially Pavlovian biases for low- and high-calorie food[14]. For each parameter, participant and population-level posterior

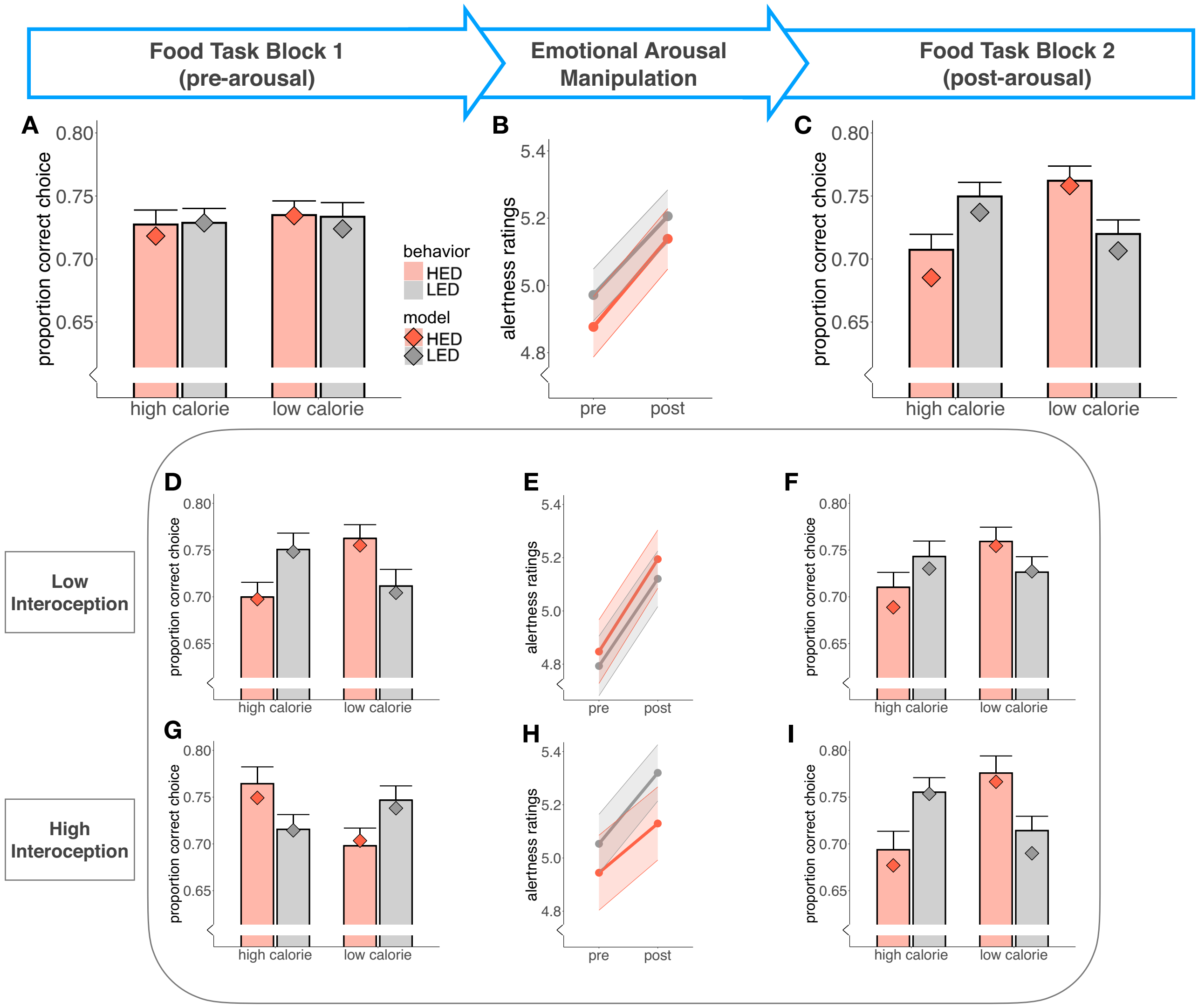

Fig. 3 | Participant- and model-generated learning during the food task in the high eating-disorder ('HED', red) and low eating-disorder ('LED', gray) symptom groups before and after the emotional arousal manipulation. A First block in the food task (pre-arousal manipulation): average proportion of correct choices in the food task when the underlying reward category was either high-calorie food stimuli ('cooked/baked' food category) or low-calorie food stimuli ('raw' food category). **B** The increase in self-reported alertness ratings (from 'sleepy' to 'alert') before ('pre') and after ('post') the wheel of fortune served as a manipulation check that the wheel of fortune generated an increase in arousal. **C** Second block in the food task (post-arousal): the emotional arousal manipulation shifted the group-specific food bias in reward learning, such that the HED group performed better when low- (versus high-) calorie stimuli were more rewarding whereas the opposite was true for the LED group. **D**–**I** Behavior in (**A**–**C**) divided by individuals with high and low interoceptive awareness. Interoceptive awareness predicted initial food biases, such that (**D**) participants with lower interoceptive awareness showed the overall group-specific food bias immediately while (**G**) those with higher interoceptive awareness demonstrated the opposite pattern. **E**, **H** Lower interoceptive awareness predicted lower overall alertness ratings. **F**, **I** The emotional arousal manipulation had a greater impact on individuals with higher interoceptive awareness by flipping their initial food biases to the overall (and replicated) pattern. A reinforcement learning model that included group-specific food biases for the initial values (pre-existing preferences for the food categories) and category-specific learning rates for positive outcomes reproduced the pattern of behavioral results. Diamond points indicate average model predictions from the best-fitting model across participants. Error bars represent the standard error of the mean.

distributions were estimated simultaneously via Hamiltonian Monte Carlo sampling[31]. WAIC scores (Watanabe-Akaike Information Criterion[32]) that penalize for model complexity were used to compare alternative models (see Methods for complete model fitting/comparison procedures and results). We further evaluated the best-fitting model through posterior predictive checks (diamond points in Figs. 2, 3) and parameter recovery (SI 4.3, Fig. S9). For computational modeling of the control task, see SI 4.1 and Fig. S7.

We first compared hierarchical models that successively included all possible combinations of each of the tested four components (see M1 to M4 in Fig. 4A for a subset of single-component model fits). We found that a model (M2 + M4) that fit distinct ('asymmetric') learning rates for positive and negative prediction errors ('positive and negative learning rates') and initial category values provided the best fit overall. We next added group-level difference parameters (HED versus LED) on the initial values and learning rates. If there were no differences between groups on these parameters, then the estimated group difference would be near 0, and the model would fit worse. We instead found that including group-level difference parameters for each learning block (pre- and post-arousal) and food category (low- versus high-calorie) in the initial category values and for positive (but not negative) learning rates, better fit behavior (Fig. 4A). Importantly, this model captured the observed patterns of learning behavior (diamond points in Figs. 2, 3) and its parameters were recoverable (SI 4.3 and

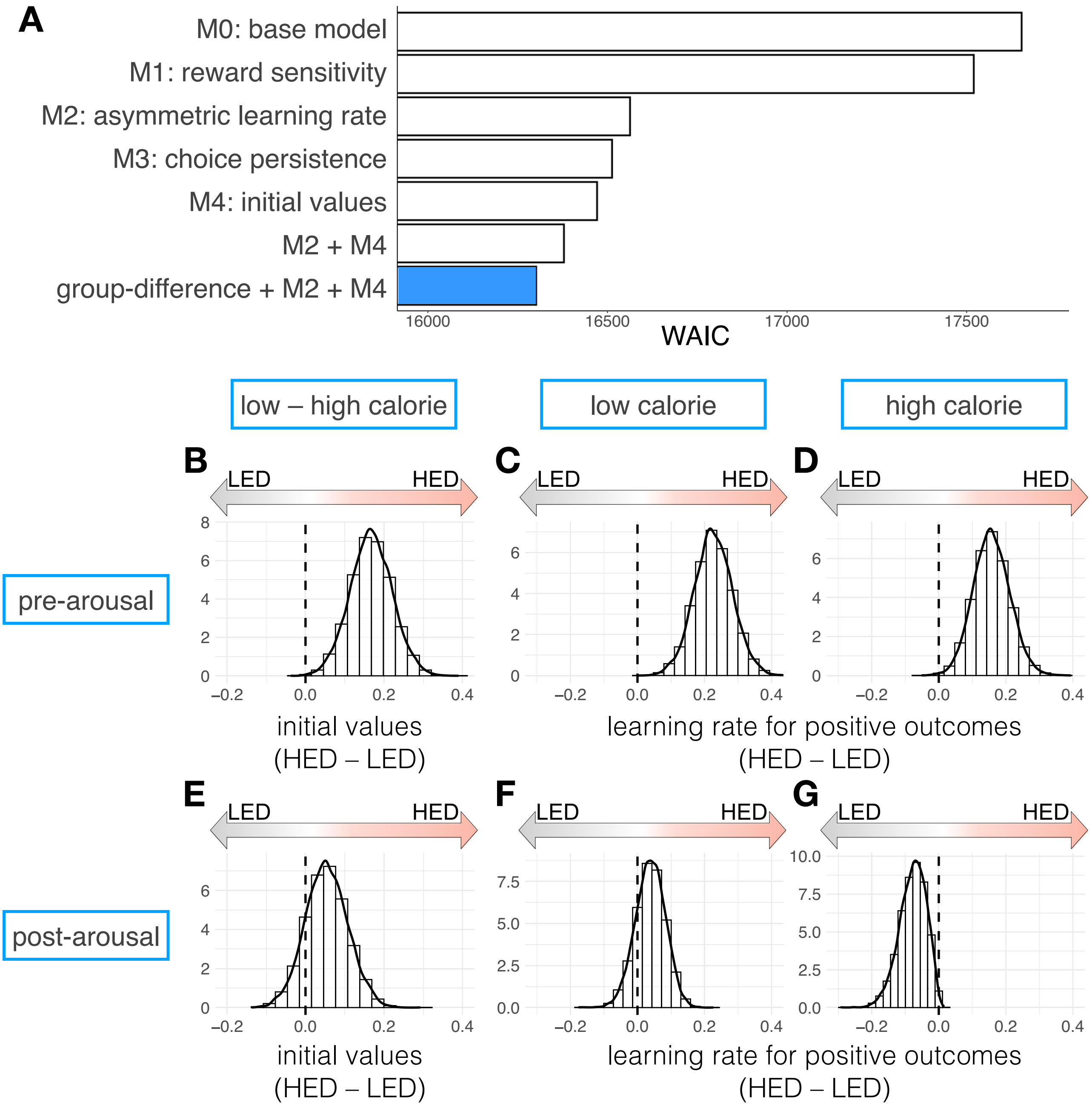

Fig. 4 | Reinforcement learning models fit to choice behavior in the food task. **A** Model comparison, lower WAIC scores indicate better model fit (the blue bar indicates the winning model). The displayed fits include the following models (see Table 1 for parameter values), 'M0': base model, standard Q-learning model; 'M1': base model adding a reward-sensitivity parameter; 'M2': base model adding different learning rates for positive and negative prediction errors; 'M3': base model adding a choice persistence parameter for a particular category; 'M4': base model with fit initial values for each food category; 'M2 + M4': model that includes the asymmetric learning rate (M2) and initial value (M4) components; 'group-difference + M2 + M4': best-fitting model that includes group-difference parameters (high eating-disorder 'HED' versus low eating-disorder 'LED' group) in the positive learning rates (M2) and in the initial values (M4) of the food categories (fit separately for each learning block). **B**–**G** Group-difference (HED versus LED) parameter distributions for the best-fitting model; y-axis indicates density. Parameters were considered significant if their 95% confidence interval did not span 0 (dotted line). In the first learning block of the food task (pre-arousal, top row), (**B**) HED had higher initial values for low- (versus high-) calorie food stimuli and (**C**, **D**) higher positive learning rates across food categories (low and high calorie). However, in the second learning block of the food task (after the emotional arousal manipulation, bottom row), group differences in (**E**) initial values for the food categories and (**F**) positive learning rate for low-calorie stimuli no longer passed significance. Nevertheless, (**G**) HED's positive learning rates for high-calorie food shifted to be lower than that estimated for LED after the emotional arousal manipulation. This change was specific to high-calorie foods (not low-calorie foods) and accounted for the behavioral pattern of food biases in the second block of the food task.

Fig. S9). We next focused on which of these parameter distributions reflected significant differences between groups (where the 95% confidence interval did not span 0; Fig. 4B–G).

During the first learning block of the food task (pre-arousal), the HED group had higher initial values for the low- (versus high-) calorie food category (Fig. 4B). This points to a pre-experimental preference in the HED group for low- versus high-calorie food (or, equivalently, an aversion against high- versus low-calorie food relative to the LED group). HED participants also demonstrated higher ('positive') learning rates for positive prediction errors in general (irrespective of high-

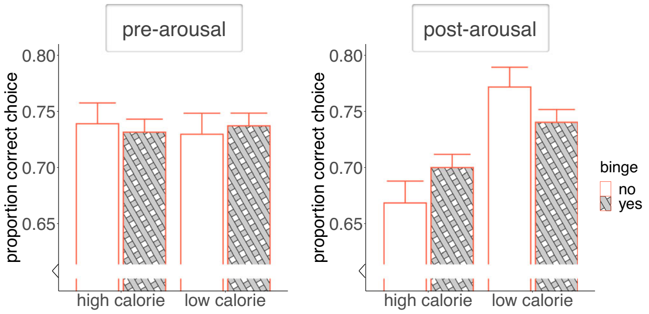

Fig. 5 | Learning performance in participants with high eating-disorder symptoms ('HED' group) who self-reported binge eating versus not. HED participants who reported binge-eating behavior in the last 6 months, relative to those who did not, showed a smaller post-arousal shift in the HED bias for low- (versus high-) calorie food. Error bars represent the standard error of the mean.

or low-calorie food category; Fig. 4C, D). Further corroborating this finding, we also found higher positive learning rates in the HED (versus LED) group before the emotional arousal manipulation in the control task (SI 4.1, Fig. S7C; no group differences in any other parameter, Fig. S7B, D). This finding indicates greater HED sensitivity (updating values more) for outcomes associated with positive prediction errors than the LED group. To note, these group differences in the first block led to similar overall behavior across groups (Fig. 3A), suggesting that interactions between initial values and learning rates produced behavioral patterns that obfuscated group differences.

After the emotional arousal manipulation in the second block of the food task, overall group-specific food biases more clearly emerged. In the model, this pattern was supported by a striking change in HED's positive learning rate for high-calorie food (but not low-calorie food), which was now significantly lower than the LED group (Fig. 4G). That is to say, individuals in the HED group were less sensitive to positive outcomes associated with high-calorie food items compared to the LED group. This relative group shift in the positive learning rate for high-calorie stimuli enabled the model to capture post-arousal food biases wherein the HED group learned to associate reward better with low- versus high-calorie food, and vice-versa for the LED group (Fig. 3C). Group differences in the initial values and positive learning rate for low-calorie stimuli were no longer significant (Fig. 4E, F). In the control task, there were no longer any group differences in fit parameters after the emotional arousal manipulation, including that for positive learning rates (Fig. S7E–G). These results suggest that the emotional arousal manipulation selectively decreased positive learning rates for high-calorie food in the HED (relative to the LED) group.

Binge-eating behavior predicted an attenuated change in food bias

Given the potential heterogeneity of our non-clinical HED sample, we evaluated individual differences in eating disorder symptoms and behaviors in an exploratory analysis (for statistical approach, see SI 1.3). While we did not find any of the tested variables to exert independent effects on learning performance, individual differences in binge-eating behavior interacted with emotional arousal effects to predict food-reward biases.

Specifically, we found that greater binge-eating behavior in HED participants (reported in the last 6 months; EAT-26 questionnaire[33]) predicted less of a post-arousal change in reward bias in the food task (mixed-effects logistic regression predicting correct choice as a function of a three-way interaction between reward category, arousal block, and the amount of binge behavior, B = −0.25, z = −3.41, $p < 0.001$; M = −0.26 [−0.41, −0.11]; Fig. 5). In other words, binge-eating behavior was associated with a relatively weaker reward bias for low-calorie food and a higher reward bias for high-calorie food compared to HED participants who did not report binge eating. This result suggests that binge eating may be less predictive of the food bias for low- versus high-calorie stimuli than other symptoms of anorexia and/or bulimia. Binge-eating disorder can moreover underlie the rapid consumption of high-calorie food[34]. While we did not explicitly recruit participants with binge-eating disorder, comorbid symptoms in our study suggest different reward associations than that for anorexia and bulimia, potentially including less avoidance of high-calorie food.

Explicit food ratings

After the reward learning task, participants explicitly evaluated a subset of high- and low-calorie food items on healthiness, liking, and calorie amount. Participants in the HED group rated high-calorie food as less healthy and low-calorie food as healthier than the LED group (mixed-effects linear regression predicting health ratings as a function of the interaction between group and food category, B = −0.27, t = −3.34, $p < 0.001$; M = −0.27 [−0.43, −0.10]). We also found that HED participants were less likely to underestimate the number of calories in high-calorie food (relative to low-calorie food) than LED participants (mixed-effects linear regression predicting calorie accuracy as a function of the interaction between group and food category, B = 116.22, t = 3.46, $p < 0.001$; M = 116.54 [49.65, 185.30]). This result is consistent with findings demonstrating greater knowledge of calorie content in eating disorders[35]. Despite these group differences in health judgment and calorie accuracy, we did not find the HED and LED groups to differ in explicit food liking: low-calorie (raw) food was liked less than high-calorie food across participants (mixed-effects linear regression predicting liking ratings as a function of food category, B = −0.37, t = −5.25, $p < 0.001$; M = −0.37 [−0.51, −0.24]), with no statistically significant difference in this effect between groups (interaction between group and food category, B = −0.02, t = −0.17, $p = 0.868$; M = −0.03 [−0.31, 0.24]).

Discussion

Food plays a pivotal role in life, influencing a wide range of behaviors including affective and cognitive processes not directly related to eating. In this study, we examined whether and how innate or learned biases for food modulate general reward learning mechanisms in participants with high ('HED') and low ('LED') eating-disorder symptoms. Given the impact of emotional arousal on dietary behaviors and goal-directed actions, we also included an emotional arousal manipulation where participants either won or lost a spin on a wheel of fortune. While participants in the LED and HED groups performed equally well in the reward learning task, we identified and replicated a food-related bias: the HED group was biased to associate low- (versus high-) calorie food with reward, performing better when the underlying reward category was represented by low-calorie stimuli ('raw' food category). In contrast, the LED group was biased to associate high- (versus low-) calorie food with reward, performing better when high-calorie stimuli ('cooked/baked' food category) represented the underlying reward category. An emotional arousal manipulation modulated these group-specific food biases regardless of valence (win or loss). Individual differences in the ability to perceive and interpret arousal—interoceptive awareness—further predicted this change. Lastly, computational modeling of choice behavior emphasized the role of pre-experimental (innate or learned) food values and learning rates for positive (prediction-error) outcomes in driving the observed food biases in reward learning.

Eating behavior holds special reigns in recruiting goal-directed and habit-learning systems as a primary reinforcer. Maladaptive eating patterns, across restricted and unrestricted eaters, have been associated with the greater use of habit-learning mechanisms in behavior and in the brain[8,12,13,36]. In the current paradigm, the task goal was to select the category associated with a higher probabilistic reward.

Images of high- and low-calorie food stimuli tested whether food preferences would elicit habit- or Pavlovian-like biases on behavior that may compete with goal-directed choice. Across our pilot study (no emotional arousal manipulation) and main study (with an emotional arousal manipulation), we replicated such food biases on reward learning. Consistent with work demonstrating attentional and motivational preference for high-calorie food in typical eaters[2,3]—a putative marker of adaptive evolution—LED participants performed better when high-calorie stimuli generated more reward. On the other hand, HED participants, potentially having learned to associate high-calorie food with punishment and low-calorie food with reward, performed better when low-calorie stimuli were more rewarding. In other words, the congruence of innate or learned reward associations with goal-directed action predicted learning performance across groups.

These food-reward associations further interfered with goal-directed behavior, indicating the potential benefit of treatment that targets habit-related, reward-learning systems. Previous interventions include goal-directed strategies, which can alter food decisions by regulating focus; for example, attending to the 'health' or 'tastiness' of food can shift choice in the goal-consistent direction[37–39], while focusing on the 'positive' or 'negative' outcomes of consuming food can increase or decrease food craving, respectively[40]. Whether these goal-directed strategies are effective for those with eating disorders is unclear. Our results encourage future studies to test interventions that target pre-experimental habits[41], which, although less flexible to change, may support long-term behavior modification. Improved decision-making in anorexia nervosa is indeed related to reduced use of habit-related regions in the brain[42].

Emotional events can further modulate internal physiological states and shift our decisions[16]. Interoception, the ability to perceive and interpret such internal signals, including hunger, satiety and general arousal, is impaired across eating disorders, potentially facilitating maladaptive eating behaviors[21,23]. In our pilot sample, we found HED participants to self-report generally lower arousal ratings, indicating potentially reduced interoception. We did not replicate this finding, possibly due to differences in the LED samples between studies. In our main study, but not the pilot study, all LED participants had self-reported at least one non-eating-related psychiatric disorder (to match the psychiatric comorbidities in our HED sample), which also predicts increased interoceptive dysfunction[43]. Instead, we found that individual differences in interoceptive awareness (across groups) predicted initial food biases as well as their change after the emotional arousal manipulation. Participants with lower interoceptive awareness immediately showed the overall group-specific food bias, which did not change after the emotional arousal manipulation. On the other hand, participants with higher interoceptive awareness initially showed a flipped learning pattern, wherein the HED group demonstrated a food bias for high- (instead of low-) calorie food and vice-versa for the LED group, indicating a reversal of potentially maladaptive (to HED or LED) food-reward associations. The emotional arousal manipulation however shifted these flipped biases to the overall replicated pattern. This underscores interoceptive awareness as a possible intervention while also highlighting the challenge posed by emotional arousal. Note that the overall pattern of group-specific food biases was immediately apparent in the pilot study, perhaps indicating that the pilot sample may have included more participants with lower interoceptive awareness. Furthermore, in an exploratory analysis, we found that certain components of interoceptive awareness[25] – related to actively engaging with and trusting physiological and emotional signals from the body—predicted this effect (but not the ability to simply notice body signals or regulate negative emotional states). Body mistrust has in fact been proposed as the core mechanism through which low interoception drives eating disorder behavior[44], including suicidal ideation[45]. Our findings suggest the possibility that interventions promoting active listening and trusting of internal, bodily signals may help reverse learned and potentially harmful food-reward associations[46]. While heightened emotional arousal may temporarily disrupt this process, we speculate that emotional arousal may also make underlying food-reward associations more labile to change during interventions. Future research should formally test such interventions.

Moreover, while group-level food biases in the main study only fully emerged after the emotional arousal manipulation, the results of computational modeling point to cognitive components that differed between groups even before emotional arousal. Capturing dynamics of trial-by-trial behavior enabled a more fine-grained and mechanistic interpretation of what, if anything, is driving different group behaviors, an approach fueling the burgeoning field of computational psychiatry[47,48]. Results of the best-fitting computational model indicated that decision-making across tasks was guided by individual variability in pre-experimental values for reward categories ('initial values') as well as the extent to which those values were updated after experiencing a rewarding or disappointing outcome ('positive and negative learning rates', respectively). Adding group-level difference parameters to the initial values and positive learning rates improved model fit in both the control and food tasks, although model fitting of the food task further benefited from group differences applied to category-specific parameters (for high- and low-calorie food).

We found higher initial values for low-calorie stimuli in the HED (compared to LED) group during the pre-arousal block, pointing to pre-experimental, potentially habit- or Pavlovian-like, reward associations with low-calorie food and/or avoidance of high-calorie food[14]. Group differences in the positive learning rates for high- and low-calorie food (fit separately), represented the extent to which a better-than-expected outcome amplified or diminished value updating of either food category (similar to a confirmation-bias model[49]). In the pre-arousal blocks, HED participants exhibited higher positive learning rates across food (and control) categories, demonstrating greater value updating or sensitivity to outcomes that were better than expected (for further discussion, see SI 4.1). We did not find group differences for negative learning rates, i.e., when outcomes disconfirmed expectations. Despite these model-estimated differences before the emotional arousal manipulation, there were no overall group differences in raw or model-generated behavior, pointing to how averaged data can mask group differences in trial-by-trial dynamics.

After the emotional arousal manipulation, there was a specific and marked group shift in the strength of positive learning rates for high-calorie stimuli. Here, HED (compared to LED) participants updated their values less for high-calorie, but not low-calorie, stimuli after experiencing outcomes that were better than expected. In other words, the emotional arousal manipulation modulated food biases by shifting learning away from positive outcomes associated with high-calorie stimuli in the HED, relative to the LED, group. No other group-difference parameters were significant after the emotional arousal manipulation (including the control task), showcasing the specificity of arousal-related effects. This finding suggests greater avoidance of high-calorie stimuli, putatively strengthened by emotional arousal, in our HED sample.

Although heightened emotional arousal can sometimes help or hurt goal-directed performance[50], we did not find our manipulation to significantly influence learning performance overall. Instead, the emotional arousal manipulation brought out more habit-like behavior in both groups, advantaging learning for stimuli consistent with pre-experimental reinforcement across individual differences (such as interoceptive awareness). We further did not find the valence of the emotion (win or loss) to significantly modulate choice behavior, suggesting global arousal effects rather than mood-congruent effects that depend on positive or negative emotion. Nevertheless, we did find that the loss condition led to a larger self-reported change in happiness ratings, indicating that a loss may have had a greater impact than a win, in line with previous literature[51]. While this difference did not significantly influence subsequent learning, it may also be because our between-subjects design for the wheel-of-fortune manipulation was underpowered to detect potential differences. Future work could moreover examine how different magnitudes and durations of negative versus positive emotion may prompt distinct food biases in typical and disordered eaters.

Given the strong negative associations with high-calorie food in anorexia and bulimia nervosa, we further investigated whether specific eating disorder behaviors may predict this avoidance. In an exploratory analysis, we examined the influence of restraint, binge, and purge behaviors in the last 6 months. Binge-eating predicted relatively less avoidance of high-calorie stimuli after the emotional arousal manipulation, which may indeed be a distinguishing factor of binge-eating disorder where food is rapidly consumed in large quantities versus anorexia and bulimia nervosa, where food is often abstained from or purged[34]. Future investigations could examine how emotional arousal, including physiological measures of arousal, may interact with food biases to predict actual eating behavior in disorders characterized by restrictive versus binge-type eating.

We lastly compared the potentially implicit food biases in our task to explicit ratings of high- and low-calorie food. We found the observed pattern reflected in judgments of health, but not liking. This finding may be inconsistent with anorectic patients' diminished perception of 'tastiness' for high-fat food[12]. Nevertheless, anorectic patients – across restricting and binge/purge subtypes[52]—make decisions based on health rather than tastiness[12,13] suggesting that the food biases observed in our task may be driven more by health associations rather than by liking. Regardless, research into the implicit and explicit preferences for high- and low-calorie food has been proposed as a promising avenue for understanding the trajectory of eating disorders and developing targeted treatment for implicit reactions[53].

While our investigation provides insight into the mechanisms supporting food-reward associations in individuals with high and low eating-disorder symptoms, it has limitations. One limitation of the current study is that our HED group was not clinically evaluated, precluding the characterization of specific disorder as well as the stage of disease or recovery. The HED samples in our pilot and main study were selected based on high EDE-Q scores (minimum of 3), indicating severe eating disorder symptoms[54]. Our main study sample was further distinguished by a self-reported history of anorexia nervosa and/or bulimia nervosa (113 out of 138 participants reported a history of anorexia nervosa). Nonetheless, our HED group represented a range of eating disorder symptoms and behaviors, which we identified and tested in an exploratory analysis. While these individual differences in our HED sample did not undermine our group results, a larger study recruiting clinically well-characterized participants for particular eating disorder behaviors may reveal relevant distinctions in food biases (such as binge-eating disorder). Additionally, our online HED sample self-reported a relatively high average weight status (as assessed by body mass index, BMI), which is unusual for anorectic patients[55]. There has however been a rise in even anorectic patients with higher BMI[56], which has called into question the utility of BMI as a marker of disease severity[57]. Higher BMI has in fact been associated with stronger eating disorder symptoms as measured by the EDE-Q[58], which we also found in our sample. To address this concern, we ensured consistency in self-reports between the screener and main study, although we acknowledge that self-reported BMI may not always be accurate[59]. We further included BMI as a nuisance variable in statistical models and did not find BMI to significantly modulate our effects of interest (all ps > 0.05; for analysis, see SI 1.4). Future studies should replicate our key findings in clinically-evaluated samples where BMI, body composition, and food consumption can be more directly assessed (controlling for self-report biases).

That being said, we did replicate an overall food bias in our HED sample across two independent studies, in line with our preregistered hypothesis. Notably, both studies varied regarding numerous factors (e.g., recruitment from different online platforms, male/female participation, emotional arousal manipulation, trial structure, etc.), demonstrating the robustness of the bias. Moreover, the results held when we controlled for and separately evaluated individual differences in self-reported eating disorder (anorexia nervosa, bulimia nervosa, or both), BMI, diagnostic history, psychotropic medication as well as specific eating disorder symptoms and behaviors. Altogether, our findings show how food-reward associations and emotional arousal may interfere with goal-directed action in people with low or high eating-disorder symptoms, highlighting potential cognitive targets for treating maladaptive eating behaviors.

Methods

Pilot study and preregistration

To address our research goals, we implemented a two-step approach. We first conducted a behavioral pilot study using our paradigm (for details, see Supplementary Information SI 5). We found female and male HED participants ($n = 108$) to show a food bias for low-calorie food, performing better when the low- versus high-calorie food category was more rewarding, whereas female and male LED participants ($n = 119$) showed the opposite bias (mixed-effects logistic regression predicting correct choice as a function of the interaction between group and reward category, B = −0.33, z = −3.71, $p < 0.001$; M = −0.33 [−0.50, −0.15]; SI 5.2 and Fig. S10A). In the pilot study, we also tested the emotional arousal manipulation (wheel of fortune) in a separate session (independent from the reward learning task, and different from our main study). HED participants self-reported overall lower arousal ratings (mixed-effects linear regression predicting ratings as a function of group, B = −0.38, t = −3.24, p = 0.00138; M = −0.37 [−0.60, −0.15]; SI 5.3 and Fig. S10C) as well as a steeper change across the emotional arousal manipulation (SI Fig. S10C), indicating potentially distinct interoceptive processing. We preregistered[60] these two findings as hypotheses to replicate in our main study. In the main study, we further examined the impact of the emotional arousal manipulation and individual interoceptive awareness on group-specific food biases.

Participants

The study protocol was approved by Caltech's Institutional Review Board (IR20-1051), and all participants provided informed consent prior to beginning. We preregistered the main study on June 10, 2022 (https://osf.io/yp6na) and report all deviations in SI 5.4. To recruit our low eating-disorder ('LED') and high eating-disorder ('HED') symptom groups, we administered a screener to 4000 participants on the online platform Prolific (half self-reported female (2000), 18–50 years old; for further detail, see SI 1). The screener included the EDE Questionnaire (EDE-Q[61]), which asked participants to rate the frequency of eating disorder symptoms and behaviors from 0 ('No days') to 6 ('Every day') as well as self-reports of eating-related and non-eating-related psychiatric disorders.

Consistent with the selection method outlined in our preregistration[60], for our HED group, we invited participants who scored at least 3—above the clinical cutoff of 2.8[54] on the global measure of the EDE-Q—and who self-reported histories of anorexia nervosa across its subtypes, including the binge/purge subtype, and bulimia nervosa, given similarities in food preference and potential diagnostic overlap[52,62]. For our LED group, we invited participants who scored less than 1.5 on the EDE-Q and did not report current or past histories of any eating disorder. Additionally, all participants must have correctly answered all attention checks, which were items randomly inserted within four questionnaires that asked participants to press a particular response[63], leading to the exclusion of 414 participants.

Based on the above criteria, we invited a total of 1003 participants to take part in our reward-learning paradigm one week later, and 575 completed our study. In line with our preregistration, we excluded participants who missed more than 5% of the decision-making trials and/or failed more than two attention checks in the questionnaires. We also excluded participants who reported an inconsistent height and weight (used to calculate BMI) between the original screener and the experiment, meaning their self-reported height could not have deviated more than one inch (in the present sample, 8% of participants reported at most a one-inch difference) and their self-reported weight could not have deviated more than 15 pounds (in the present sample, 90% of participants reported a fluctuation of less than five pounds).

While we initially recruited an even number of participants based on sex assigned at birth, the majority of participants who qualified for the eating disorder group were female (138 out of 165). Due to this low sample of male participants in the HED group (statistically underpowered to detect effects, see SI 1.1), we excluded male participants from both groups. We further matched the HED group ($n = 138$) with participants from the LED group ($n = 141$) based on age, education, and self-report of at least one non-eating-related psychiatric disorder (given comorbidities in eating disorders). We did not find any non-eating-related psychiatric disorder to significantly predict group-specific food biases (SI 1). We were not able to match on BMI (BMI was tested and controlled for across all analyses with no statistically significant effects, see SI 1.4). For characteristics of the final sample, see SI 1 and Table S1.

While our heterogeneous HED sample is defined by disordered eating rather than clinically-diagnosed eating disorders, self-reported histories of anorexia ($n = 113$) or bulimia ($n = 84$; 59 indicating both) along with active eating disorder symptoms, suggest that HED participants may have developed food-reward associations similar to those observed in clinical populations. Moreover, the heterogeneity of a large online sample enabled us to examine whether specific eating disorder symptoms and behaviors predicted task outcomes (SI 1.3). We checked for differences in behavior between self-reported eating disorder, diagnostic history, current psychotropic medication, as well as current eating disorder behaviors (e.g., restraint, binge, and purge behaviors; for HED sample characteristics and analysis, see SI 1.3). These exploratory analyses did not find any variable related to eating disorder symptoms and behaviors to exert significant, independent effects on learning performance, nor did any subgroup exhibit a different pattern of results. However, in supplementary post-hoc analyses, we found individual differences in binge eating to interact with the change in food bias across the emotional arousal manipulation, which we report as an exploratory finding in the manuscript.

Procedure

Participants in the main study completed three sections: (1) reward-learning task (Figs. 1), (2) eight short questionnaires evaluating eating behaviors and mental health (SI 1.2), and (3) assessment of a selection of food stimuli on health, liking, and calorie accuracy (participants were not aware of these assessments when they completed the reward-learning task). The online study took, on average, 40 min to complete, and participants were paid a rate of $15 per hour.

Reward learning task

The probabilistic learning task consisted of six blocks in the following order: (1) practice block, (2–3) two learning blocks (one of the 'food task' and one of the 'control task'), (4) wheel-of-fortune emotional arousal manipulation (win or loss), and (5–6) two post-manipulation learning blocks (one per food and control task; Fig. 1B). Across participants, we counterbalanced three task elements: (1) the order of the food and control tasks, (2) whether the wheel of fortune resulted in a win or a loss, and (3) which stimulus category was initially more rewarding in each block, leading to a total of 8 configurations. The trial-unique food and scene stimuli were randomly selected on each trial (each item only appeared once in the reward-learning task; for stimuli information, see SI 2). Food stimuli were selected from the 'Food-Pics_Extended' database[64] and the scene stimuli were selected from the SUN database[65] (SI 2).

Participants were instructed to learn which of two categories on a trial (represented by two stimuli displayed on the screen) was more rewarding through trial and error; they chose one stimulus and received feedback about the outcome (rewarded or not). During the practice block (14 trials), participants were walked through the task, made decisions between indoor and outdoor scenes (control task), and experienced one 'change-point' event to the underlying reward distribution. A change-point event indicated a reversal of the rewards associated with each category, meaning the more rewarding category became the less rewarding one. They had to correctly answer all comprehension questions before moving on to the actual task.

Each learning block included either two food items as categories ('food task') or two scene categories ('control task'; Fig. 1A). One stimulus category was always more rewarding than the other (75% versus 25% reward probability). The categories were specified in the instructions but were not labeled on decision-making trials. In the food task, the categories were raw versus cooked/baked foods, which highly correlated with the calorie content of the food and served as a cover story for low- and high-calorie foods, respectively (high-calorie mean calories: 1208.14; low-calorie mean calories: 99.18). The categories in the control task were indoor and outdoor scenes. Each learning block consisted of 40 trials, and on trial 20, the reward categories reversed (change-point). On each trial, participants had 5 s to make a choice ('1') for the left category and '2' for the right category, or else they would miss the outcome on that trial; category placement (left or right side of the screen) was randomized. On average, participants missed less than 0.2% of trials. The outcome of any trial was presented for 1 s and showed either '$0' ('not rewarded') or '$10' ('rewarded'), a small proportion of which participants were told they would receive (in reality, participants received the same payment of $15/hour for this task regardless of performance). Each trial was separated by a one-second inter-trial-interval.

After two learning blocks, participants 'spun' a digital wheel of fortune via a key press (emotional arousal manipulation). They were told that the length of their press predicted the wheel's outcome; however, the outcome of the wheel (winning or losing $8) was predetermined. Participants were told the full $8 would be added or removed from their earnings.

Additionally, throughout the learning task, participants rated their current state on four measures: arousal (denoted as 'alertness' to avoid semantic confusion and positive/negative connotations), happiness, hunger, and thirst on Likert scales ranging from 1 through 6 (see SI 3.2 for further detail). Alertness and happiness ratings were collected to verify changes induced by the emotional arousal manipulation (see Results). We also checked for changes in thirst and hunger ratings, which did not significantly change across task (see SI 3.2 and Fig. S5 for visualization and analysis).

Food-stimuli ratings

After completing the questionnaires, participants rated a selection of high and low-calorie food items, 20 from each category (40 in total). They rated each food item (randomized order) on the perceived healthiness on a 6-point scale ranging from 1 'very unhealthy' to 6 'very healthy', and on how much they liked each food on a 6-point scale ranging from 1 'dislike very much' to 6 'like very much'. Participants had a maximum of 10 s to answer. They next estimated the total number of calories for 20 food items (10 low and 10 high-calorie items) and had a maximum of 15 s to answer.

Table 1 | Parameters for base and best-fitting reinforcement learning models in the food task

	learning rate (α)	beta (β)	# added parameters	added parameter values
M0: base (Rescorla-Wagner)	0.44	3.04	baseline	
M1: reward sensitivity	0.43	3.39	1	ρ = 0.96
M2: asymmetric learning rate	α + = 0.81 α- = 0.33	2.91	1	(see α)
M3: choice persistence	0.56	2.22	1	φ = 0.44
M4: initial values	0.56	2.36	1	$Q_0 = -0.85$
asymmetric learning rate (M2) + initial values (M4)	α + = 0.76 α- = 0.50	2.33	2	(see α) $Q_0 = -0.67$
asymmetric learning rate (M2) + initial values (M4) + group differences (by block)	α + = 0.72 α- = 0.50	2.37	8	(see α) $Q_0 = -0.67$ (see Fig. 3 for difference parameters)

Learning rate, beta and added parameter values indicate the means of the posterior distributions for that parameter in the corresponding model.

Statistical approach

For repeated-measures statistical tests (all two-tailed), we used mixed-effects models (R package, lme4[66]) and Bayesian multilevel modeling (R package, brms[67]) to generate estimates and confidence intervals for reported effects; M denotes the mean of the posterior distribution and the bracketed range denotes the 95% Credibility Interval (CI). We considered an effect reliable if the Bayesian CI did not span 0. We treated participants as a random effect for both the slope and the intercept of each fixed effect. If the model did not converge due to insufficient data in estimating parameters, we incrementally simplified the random effect structure until convergence was achieved (the simplest structure included participant intercept as a random effect). To predict binary choice data in the reward-learning task, we ran generalized linear mixed-effects logistic regression models (pooling across reward change-points) and otherwise ran linear mixed-effects regression models. Data met the assumptions of each statistical test. In each model, we also statistically accounted for individuals' BMI as a nuisance variable, although we never found BMI to exert statistically significant effects (see SI 1.4 for further detail).

Computational modeling

Model specifications. Our baseline model was a standard Q-learning model[26], where the value ('Q') of a chosen stimulus category ('s') is updated during a trial (t) using the following equation,

$$Q^s_{t+1} = Q^s_t + \alpha\left(r_t - Q^s_t\right) \tag{1}$$

a fit learning rate (α) determines the extent to which the reward prediction error on that trial (calculated by subtracting the value of that category, Q^s_t, from the reward outcome, r_t) is used to change the value of that category on the next trial. While outcomes were either a win ($10) or not a win ($0), we coded r_t as 1 and −1, respectively. To note, coding the reward omission as 0 rather than −1 is functionally equivalent as the decision function relies on relative rather than absolute values.

We modeled the probability of choosing a stimulus category during a trial (p^s_t) using a softmax function,

$$p^s_t = \frac{\exp(\beta Q^s_t)}{\sum^s_{i=1}\exp(\beta Q^i_t)} \tag{2}$$

where the fit inverse temperature (β) governs the stochasticity in choice.

We next investigated four potential processes that could support participant-level biases within this standard reinforcement-learning model:

(M1) Reward sensitivity ('rs'): We implemented a model where the reward outcome was scaled by a parameter (ρ), capturing potential differences in outcome sensitivity:

$$r_t = \rho * r_t \tag{3}$$

(M2) Asymmetric learning-rate: We tested whether positive and negative reward prediction errors may have an asymmetric effect on learning rate. In this model, separate learning rates were fit when the reward prediction error was positive (α^+) and negative (α^-):

$$\alpha_t = \begin{cases} \alpha^+, \text{if}(r_t - Q^s_t) \geq 0, \\ \alpha^-, \text{if}(r_t - Q^s_t) < 0. \end{cases} \tag{4}$$

(M3) Choice persistence ('cp'): In order to model the tendency for participants to repeatedly choose a category (regardless of outcomes), we fit a persistence model which bonused the value of a category (ϕ) if it had been chosen one trial back:

$$Q^s_t = \begin{cases} Q^s_{t-1} + \phi, \text{if chosen } t-1, \\ Q^s_{t-1}, \text{if not chosen } t-1. \end{cases} \tag{5}$$

(M4) Initial values: Given that preferences for high and low-calorie food were established prior to this task, we included a model that fit category values (Q_0) for the first trial of each block:

$$Q^s_t = \begin{cases} Q_0, \text{if } t = 1, \\ Q^s_t, \text{if } t > 1. \end{cases} \tag{6}$$

We tested each of these four components independently on learning across the emotional arousal manipulation (food and control task were modeled separately; food task: Table 1, Fig. 4A; control task: SI 4.1, Fig. S7), and then ran models that included different combinations of each (see 'Model fitting and comparison', below). We found that a model including (M2) an asymmetric learning rate for positive and negative prediction errors and (M4) initial values best fit choice data overall.

We next investigated whether group-level difference parameters (HED versus LED) for the components of the model (initial category values and learning rates for positive prediction errors) improved fit. In the control task, group-difference parameters fit across the scene categories (i.e., no group-specific scene bias) provided a better fit (model predictions in Fig. 2; model fits in Fig. S7). In the food task, we further found that adding category-specific group-difference parameters (for the high- and low-calorie food categories) better fit choice data and predicted biased patterns of learning (model predictions in Figs. 2, 3, model fits in Fig. 4A). Note that without comparing groups, choice persistence may be equivalent to asymmetric learning rates in this behavioral context[68]. However, allowing single parameters to change in the asymmetric learning model captured group differences better than choice persistence.

Model fitting and comparison

We fit hierarchical reinforcement learning models to choice data using Stan[31], modeling the food and control tasks separately. Parameters from Eqs. 1–6 were simultaneously fit at both the participant and (across-group) population level using Hamiltonian Monte Carlo sampling. Population-level parameters were drawn from normal distributions, and participant-level parameters were drawn from uniform distributions that placed equal probability on all possible values (through the Phi_approx function[69]). We report the population-level (hyperparameter) distribution means of fit parameters in Table 1 and include the distributions in SI Fig. S8. We ran each model 2000 iterations (500 warmup) on 4 chains, and verified that there were no divergences, a BFMI (Bayesian Fraction of Missing Information) of ≥0.2 for all chains, R-hat values ≤ 1.01, and effective sample sizes ≥400 for all parameters[70]. If the number of effective sample sizes was too low, we ran the model for 5000 iterations. All included models met these diagnostic criteria. RStan scripts are publicly available: https://zenodo.org/records/14676167.

To compare models, we computed the Widely Applicable (or Watanabe-Akaike) Information Criterion ('WAIC'[32]) for each model (Fig. 4A). The WAIC reflects a model's out-of-sample predictive performance by measuring the difference between the average log-likelihood of a model for the observed data with the variance of the average log-likelihood for simulated data drawn from the posterior distribution. This approach is considered robust to overfitting as it adjusts for the effective number of parameters in the model, accounting for model complexity. Lower scores indicate a better fit.

We conducted checks on the best-fitting model, including posterior predictive checks (model predictions in Figs. 2 and 3) and parameter recovery, see SI 4.3 and Fig. S9.

Group-difference parameters from best-fitting model

We added group-difference parameters to the two components of the model (initial category values and positive learning rate) on each learning block (pre- and post-arousal); food task: Fig. 4B–G, control task: Fig. S7B–G. For the food task, group-difference parameters were fit on low (versus high) initial Q-values (2 parameters in total, 1 for each block), and were fit separately on positive learning rates for high- and low-calorie food stimuli (4 parameters in total, 2 for each block). Note that fitting group-difference parameters on negative learning rates did not improve fit and were not included in the winning model. For the control task, group-difference parameters were fit on overall (rather than category-specific) parameters, as we did not expect nor find group biases for scene categories (for model fitting and comparison to the control task, see SI 4.1 and Fig. S7). We considered parameter distributions to reflect significant differences between the HED and LED groups only if their 95% confidence interval did not span 0.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All (de-identified) data used in this study are available[71]: https://zenodo.org/records/14676167.

Code availability

All code (and data) to reproduce statistical results and figures are available[71]: https://zenodo.org/records/14676167.

References

1. Rossi, M. A. & Stuber, G. D. Overlapping brain circuits for homeostatic and hedonic feeding. *Cell Metab.* **27**, 42–56 (2018).
2. Cunningham, C. A. & Egeth, H. E. The capture of attention by entirely irrelevant pictures of calorie-dense foods. *Psychon. Bull. Rev.* **25**, 586–595 (2018).
3. Guleken, Z. & Uzbay, T. Neurobiological and neuropharmacological aspects of food addiction. *Neurosci. Biobehav. Rev.* **139**, 104760 (2022).
4. Arcelus, J., Mitchell, A. J., Wales, J. & Nielsen, S. Mortality rates in patients with anorexia nervosa and other eating disorders. A meta-analysis of 36 studies. *Arch. Gen. Psychiatry* **68**, 724–731 (2011).
5. Silén, Y. & Keski-Rahkonen, A. Worldwide prevalence of DSM-5 eating disorders among young people. *Curr. Opin. Psychiatry* **35**, 362–371 (2022).
6. Monteleone, A. M. et al. Treatment of eating disorders: A systematic meta-review of meta-analyses and network meta-analyses. *Neurosci. Biobehav. Rev.* **142**, 104857 (2022).
7. Brown, R. M., Dayas, C. V., James, M. H. & Smith, R. J. New directions in modelling dysregulated reward seeking for food and drugs. *Neurosci. Biobehav. Rev.* **132**, 1037–1048 (2022).
8. Koban, L., Wager, T. D. & Kober, H. A neuromarker for drug and food craving distinguishes drug users from non-users. *Nat. Neurosci.* **26**, 316–325 (2023).
9. O'Hara, C. B., Campbell, I. C. & Schmidt, U. A reward-centred model of anorexia nervosa: a focussed narrative review of the neurological and psychophysiological literature. *Neurosci. Biobehav. Rev.* **52**, 131–152 (2015).
10. Steinglass, J. E. & Foerde, K. How does anorexia nervosa become resistant to change? in *Managing severe and enduring anorexia nervosa* 64–75 (Routledge, 2016).
11. Foerde, K. et al. Deficient goal-directed control in a population characterized by extreme goal pursuit. *J. Cogn. Neurosci.* **33,** 463–481 (2021).
12. Foerde, K. et al. Restrictive eating across a spectrum from healthy to unhealthy: behavioral and neural mechanisms. *Psychol. Med.* **52**, 1755–1764 (2022).
13. Foerde, K., Steinglass, J. E., Shohamy, D. & Walsh, B. T. Neural mechanisms supporting maladaptive food choices in anorexia nervosa. *Nat. Neurosci.* **18**, 1571–1573 (2015).
14. Vogel, V. et al. Pavlovian-to-instrumental transfer in Anorexia Nervosa: A pilot study on conditioned learning and instrumental responding to low- and high-calorie food stimuli. *Eur. J. Neurosci.* **51**, 1794–1805 (2020).
15. Cowdrey, F. A., Finlayson, G. & Park, R. J. Liking compared with wanting for high- and low-calorie foods in anorexia nervosa: aberrant food reward even after weight restoration. *Am. J. Clin. Nutr.* **97**, 463–470 (2013).
16. Poppa, T. & Bechara, A. The somatic marker hypothesis: revisiting the role of the 'body-loop' in decision-making. *Curr. Opin. Behav. Sci.* **19**, 61–66 (2018).
17. Connan, F., Campbell, I. C., Katzman, M., Lightman, S. L. & Treasure, J. A neurodevelopmental model for anorexia nervosa. *Physiol. Behav.* **79**, 13–24 (2003).
18. Evers, C., Dingemans, A., Junghans, A. F. & Boevé, A. Feeling bad or feeling good, does emotion affect your consumption of food? A meta-analysis of the experimental evidence. *Neurosci. Biobehav. Rev.* **92**, 195–208 (2018).
19. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as representation of momentum. *Trends Cognitive Sci.* 20 15–24 https://doi.org/10.1016/j.tics.2015.07.010 (2016).
20. Critchley, H. D. & Garfinkel, S. N. Interoception and emotion. *Curr. Opin. Psychol.* **17**, 7–14 (2017).
21. Martin, E., Dourish, C. T., Rotshtein, P., Spetter, M. S. & Higgs, S. Interoception and disordered eating: A systematic review. *Neurosci. Biobehav. Rev.* **107**, 166–191 (2019).
22. De la Cruz, F. et al. Central autonomic network alterations in anorexia nervosa following peripheral adrenergic stimulation. *Biol.*

Psychiatry: Cognitive Neurosci. Neuroimaging https://doi.org/10.1016/j.bpsc.2022.12.009 (2022).

23. Jenkinson, P. M., Taylor, L. & Laws, K. R. Self-reported interoceptive deficits in eating disorders: A meta-analysis of studies using the eating disorder inventory. *J. Psychosom. Res.* **110**, 38–45 (2018).
24. Eldar, E. & Niv, Y. Interaction between emotional state and learning underlies mood instability. *Nat. Commun.* **6**, 6149 (2015).
25. Mehling, W. E., Acree, M., Stewart, A., Silas, J. & Jones, A. The multidimensional assessment of interoceptive awareness, version 2 (MAIA-2). *PLoS One* **13**, e0208034 (2018).
26. Watkins, C. J. C. H. & Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
27. Huys, Q. J., Pizzagalli, D. A., Bogdan, R. & Dayan, P. Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biol. Mood Anxiety Disord.* **3**, 12 (2013).
28. Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T. & Hutchison, K. E. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl Acad. Sci. Usa.* **104**, 16311–16316 (2007).
29. Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32**, 551–562 (2012).
30. Eckstein, M. K. et al. The interpretation of computational model parameters depends on the context. *Elife* **11**, e75474 (2022).
31. Carpenter, B. et al. Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
32. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594 (2010).
33. Garner, D. M., Olmsted, M. P., Bohr, Y. & Garfinkel, P. E. The eating attitudes test: psychometric features and clinical correlates. *Psychol. Med.* **12**, 871–878 (1982).
34. Davis, H. A., Graham, A. K. & Wildes, J. E. Overview of binge eating disorder. *Curr. Cardiovasc. Risk Rep.* **14**, 1–10 (2020).
35. Simpson, C. C. & Mazzeo, S. E. Calorie counting and fitness tracking technology: Associations with eating disorder symptomatology. *Eat. Behav.* **26**, 89–92 (2017).
36. Xue, A. M. et al. Neural representations of food-related attributes in the human orbitofrontal cortex during choice deliberation in anorexia nervosa. *J. Neurosci.* **42**, 109–120 (2022).
37. Tusche, A. & Hutcherson, C. A. Cognitive regulation alters social and dietary choice by changing attribute representations in domain-general and domain-specific brain circuits. *Elife* **7**, e31185 (2018).
38. Hutcherson, C. A. & Tusche, A. Evidence accumulation, not "self-control", explains dorsolateral prefrontal activation during normative choice. *Elife* **11**, e65661 (2022).
39. Janet, R. et al. Body mass index-dependent shifts along large-scale gradients in human cortical organization explain dietary regulatory success. *Proc. Natl Acad. Sci. Usa.* **121**, e2314224121 (2024).
40. Sun, W. & Kober, H. Regulating food craving: From mechanisms to interventions. *Physiol. Behav.* **222**, 112878 (2020).
41. Steinglass, J. E. et al. Targeting habits in anorexia nervosa: a proof-of-concept randomized trial. *Psychol. Med.* **48**, 2584–2591 (2018).
42. Foerde, K. et al. Changes in brain and behavior during food-based decision-making following treatment of anorexia nervosa. *J. Eat. Disord.* **9**, 48 (2021).
43. Khalsa, S. S. et al. Interoception and mental health: A roadmap. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 501–513 (2018).
44. Brown, T. A. et al. Body mistrust bridges interoceptive awareness and eating disorder symptoms. *J. Abnorm. Psychol.* **129**, 445–456 (2020).
45. Perry, T. R., Wierenga, C. E., Kaye, W. H. & Brown, T. A. Interoceptive awareness and suicidal ideation in a clinical eating disorder sample: The role of body trust. *Behav. Ther.* **52**, 1105–1113 (2021).
46. Boswell, J. F., Anderson, L. M. & Anderson, D. A. Integration of interoceptive exposure in eating disorder treatment. *Clin. Psychol. (N. Y.)* **22**, 194–210 (2015).
47. Haynos, A. F., Widge, A. S., Anderson, L. M. & Redish, A. D. Beyond description and deficits: How computational psychiatry can enhance an understanding of decision-making in anorexia nervosa. *Curr. Psychiatry Rep.* **24**, 77–87 (2022).
48. Radzikowska, M., Pike, A. C. & Hall-McMaster, S. Computational perspectives on behaviour in anorexia nervosa: A systematic review. https://doi.org/10.31234/osf.io/wq4cz (2023).
49. Doll, B. B., Hutchison, K. E. & Frank, M. J. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J. Neurosci.* **31**, 6188–6198 (2011).
50. Yerkes, R. M. & Dodson, J. D. The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* **18**, 459–482 (1908).
51. Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263 (1979).
52. Uniacke, B. et al. A comparison of food-based decision-making between restricting and binge-eating/purging subtypes of anorexia nervosa. *Int. J. Eat. Disord.* **53**, 1751–1756 (2020).
53. Paslakis, G., Scholz-Hehn, A. D., Sommer, L. M. & Kühn, S. Implicit bias to food and body cues in eating disorders: a systematic review. *Eat. Weight Disord.* **26**, 1303–1321 (2021).
54. Velkoff, E. A., Brown, T. A., Kaye, W. H. & Wierenga, C. E. Using clinical cutoff scores on the eating disorder examination-questionnaire to evaluate eating disorder symptoms during and after naturalistic intensive treatment. *Eat. Disord.* **31**, 464–478 (2023).
55. Redgrave, G. W. et al. Refeeding and weight restoration outcomes in anorexia nervosa: Challenging current guidelines. *Int. J. Eat. Disord.* **48**, 866–873 (2015).
56. Garber, A. K. et al. Weight loss and illness severity in adolescents with atypical anorexia nervosa. *Pediatrics* **144**, e20192339 (2019).
57. Ramaswamy, N. & Ramaswamy, N. Overreliance on BMI and delayed care for patients with higher BMI and disordered eating. *AMA J. Ethics* **25**, E540–E544 (2023).
58. Ralph-Nearman, C., Yeh, H.-W., Khalsa, S. S., Feusner, J. D. & Filik, R. What is the relationship between body mass index and eating disorder symptomatology in professional female fashion models? *Psychiatry Res* **293**, 113358 (2020).
59. Gosse, M. A. How accurate is self-reported BMI? *Nutr. Bull.* **39**, 105–114 (2014).
60. Rouhani, N., Adolphs, R. & Tusche, A. Decision-making in eating disorders. Preregistration: https://osf.io/yp6na (2022).
61. Mond, J. M., Hay, P. J., Rodgers, B., Owen, C. & Beumont, P. J. V. Validity of the Eating Disorder Examination Questionnaire (EDE-Q) in screening for eating disorders in community samples. *Behav. Res. Ther.* **42**, 551–567 (2004).
62. Stoner, S. A., Fedoroff, I. C., Andersen, A. E. & Rolls, B. J. Food preferences and desire to eat in anorexia and bulimia nervosa. *Int. J. Eat. Disord.* **19**, 13–22 (1996).
63. Zorowitz, S., Solis, J., Niv, Y. & Bennett, D. Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nat. Hum. Behav.* **7**, 1667–1681 (2023).
64. Blechert, J., Lender, A., Polk, S., Busch, N. A. & Ohla, K. Food-Pics_Extended-An image database for experimental research on eating and appetite: Additional images, normative ratings and an updated review. *Front. Psychol.* **10**, 307 (2019).
65. Xiao, J., Hays, J., Ehinger, K. A., Oliva, A. & Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 3485–3492 (2010).
66. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw., Artic.* **67**, 1–48 (2015).

67. Bürkner, P.- C. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).
68. Sugawara, M. & Katahira, K. Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Sci. Rep.* **11**, 3574 (2021).
69. Stan Development Team. RStan: the R interface to Stan. http://mc-stan.org/ (2020).
70. Baribault, B. & Collins, A. G. E. Troubleshooting Bayesian cognitive models. *Psychol. Methods* https://doi.org/10.1037/met0000554 (2023).
71. Rouhani, N. *Eating Disorder Symptoms and Emotional Arousal Modulate Food Biases during Reward Learning*. (Zenodo, 2025). https://doi.org/10.5281/zenodo.14676167.

Acknowledgements

We thank Ralph Adolphs and Cindy Hagan for their suggestions and support of this project. This work was supported by the Davis Senior Postdoctoral Fellowship (The Hilda and Preston Davis Foundation Award Programs in Eating Disorders Research) awarded to A.T. and N.R.

Author contributions

N.R. and A.T. conceived the experiment. N.R. developed the task and acquired the data. N.R., A.T., and C.D.G. conceived the analyses. N.R. analyzed the data and wrote the manuscript with input from A.T., C.D.G. and J.F.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57872-w.

Correspondence and requests for materials should be addressed to Nina Rouhani.

Peer review information *Nature Communications* thanks Nils Kroemer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.