The effects of induced positive and negative affect on Pavlovian-instrumental interactions

Isla Weber[1], Sam Zorowitz[1], Yael Niv[1,2], Daniel Bennett[3]

1 Princeton Neuroscience Institute, Princeton University, USA

2 Department of Psychology, Princeton University, USA

3 School of Psychological Sciences, Monash University, Australia

Abstract

Across species, animals have an intrinsic drive to approach appetitive stimuli and to withdraw from aversive stimuli. In affective science, influential theories of emotion link positive affect with strengthened behavioral approach and negative affect with avoidance. Based on these theories, we predicted that individuals' positive and negative affect levels should particularly influence their behavior when innate Pavlovian approach/avoidance tendencies conflict with learned instrumental behaviors. Here, across two experiments—exploratory Experiment 1 (N = 91) and a preregistered confirmatory Experiment 2 (N = 335)—we assessed how induced positive and negative affect influenced Pavlovian-instrumental interactions in a reward/punishment Go/No-Go task. Contrary to our hypotheses, we found no evidence for a main effect of positive/negative affect on either approach/avoidance behavior or Pavlovian-instrumental interactions. However, we did find evidence that the effects of induced affect on behavior were moderated by individual differences in self-reported behavioral inhibition and gender. Exploratory computational modelling analyses explained these demographic moderating effects as arising from positive correlations between demographic factors and individual differences in the strength of Pavlovian-instrumental interactions. These findings serve to sharpen our understanding of the effects of positive and negative affect on instrumental behavior.

Keywords

reinforcement learning; experimental affect induction; computational modelling; Pavlovian learning; instrumental learning

Introduction

Humans and other animals have an innate Pavlovian tendency to approach stimuli that are associated with appetitive outcomes and to withdraw from stimuli associated with aversive outcomes (Brown, 1948; Carver & White, 1994; Chen & Bargh, 1999; Eliot, 2008). This distinction between approach motivation and avoidance motivation is fundamental to many theories of instrumental behavior (e.g., Dickinson & Dearing, 1979; Gray, 1975; Higgins, 1997; Konorski, 1967). Separately, the approach/avoidance dichotomy also underpins several influential accounts of human affect and emotion (Cacioppo et al., 1999; Lang, 1995; Watson et al., 1999) that posit that positive affective states potentiate approach behavior and negative affect potentiates avoidance behavior.

In some circumstances, Pavlovian approach/avoidance tendencies may interfere with learning of instrumental behavior (Chen & Bargh, 1999; Dayan et al., 2006; Dayan & Balleine, 2002; Hershberger, 1986). For instance, using a task in which cards printed with pleasant or unpleasant words (e.g., 'tasty', 'putrid') were mounted on a conveyor belt, Solarz (1960) showed that human participants were faster and less error-prone when learning behavioral responses that were congruent with stimulus valence (bringing pleasant words towards oneself, sending unpleasant words away) compared with learning of stimulus-incongruent behavioral responses (sending pleasant words away, bringing unpleasant words closer). More recently, research using the reward/punishment Go/No-Go task (Guitart-Masip et al., 2011) has confirmed that human participants readily learn instrumental responses that are congruent with their Pavlovian response biases (i.e., active 'Go' response for acquiring reward; inhibitory 'No-Go' response to avoid punishment), but show slower learning of Pavlovian-incongruent instrumental responses (Cavanagh et al., 2013; Crockett et al., 2009; Csifcsák et al., 2020; Dorfman & Gershman, 2019;

Geurts et al., 2013; Guitart-Masip et al., 2011, 2012; Millner et al., 2018; Moutoussis et al., 2018; Raab & Hartley, 2020; Swart et al., 2017, 2018). We refer to this effect as a Pavlovian-instrumental interaction, because the rates at which participants perform instrumental behaviors (i.e., 'Go' or 'No-Go' responses) vary depending on the presence of cues that have Pavlovian associations with gain versus loss.

Despite robust overall effects at the group level, however, there is marked inter-individual variability in the strength of Pavlovian influence on instrumental learning (Albrecht et al., 2016; Dorfman & Gershman, 2019; Mkrtchian et al., 2017; Moutoussis et al., 2018; Raab & Hartley, 2020). Individual differences in the strength of Pavlovian influence on instrumental learning covary with developmental stage (Raab & Hartley, 2020), with IQ scores (Moutoussis et al., 2018), and with the strength of frontal neural oscillations in the theta frequency band (Cavanagh et al., 2013); as such, it appears unlikely that these individual differences solely reflect noise or measurement error. Of particular note, a recent longitudinal study found that individual differences in Pavlovian-instrumental interaction strength had relatively weak test-retest reliability after a delay of approximately 18 months in a sample of adolescents and young adults (Moutoussis et al., 2018). This result suggests that interindividual differences in strength of Pavlovian-instrumental interactions might be the result of transient state differences between participants in addition to any stable trait differences.

One potential source of individual differences in Pavlovian-instrumental interaction strength is participants' *affective state* during behavioral testing. That is, some theories of emotion propose that positive and negative affective states can be broadly distinguished by their differing motivational tendencies (for review see Elliot et al., 2013), with positive affect associated with a potentiation of approach behaviors and negative affect associated with

potentiation of behavioral inhibition and avoidance (Cacioppo et al., 1999; Lang, 1995; Watson et al., 1999). If so, we might expect individual differences in participants' levels of (state) positive and negative affect while completing a learning task to influence the strength of their approach and avoidance tendencies, potentially manifesting as systematic differences in the strength of Pavlovian-instrumental interactions across participants.

However, the mechanism by which affective states might influence Pavlovian-instrumental interactions is not clear from the theoretical literature. Some theories (e.g., Cacioppo et al., 1999) suggest that positive affect is linked with potentiated approach behavior for appetitive stimuli specifically (but not necessarily for aversive stimuli), and that negative affect produces increased avoidance of aversive stimuli (but not necessarily of appetitive stimuli). Other theories suggest that positive and negative affect are respectively associated with a *generalized* potentiation of approach and avoidance behavior, not linked to specific stimuli. This has been shown, for instance, in the domain of reflex control, where the strength of reflexive responses to auditory startle probes is modulated by the affective valence of concurrently viewed images (Bradley et al., 1990; Lang et al., 1990; Vrana et al., 1988). Similarly, the phenomenon of conditioned suppression in animal learning has been interpreted as suggesting that a negative affective state induced by a conditioned punisher can inhibit behavioral approach to rewarding stimuli (McNaughton, 1989).

The present study sought to shed light on this question by testing to what extent experimentally induced positive and negative affect modulated the influence of concurrently presented appetitive or aversive Pavlovian cues on instrumental learning of approach/avoidance behavior. Given that the literature does not make a clear *a priori* prediction regarding the exact nature of the effects of positive/negative affect on such Pavlovian-instrumental interactions, we

adopted a two-stage exploration/confirmation study design. As such, we conducted two experiments—an exploratory Experiment 1 and a preregistered replication in Experiment 2—in which participants completing a standard reward/punishment Go/No-Go task were presented with concurrent video-based positive or negative affect inductions. We used computational modelling of behavior to quantify the strength of Pavlovian influences on instrumental learning, and to determine how task behavior might be modulated by transient differences in state positive/negative affect.

More broadly, generalized potentiation of approach (avoidance) motivation by positive (negative) affect would also be in keeping with psychiatric theories of major depression and bipolar disorder that link mania (i.e., extreme positive affect) with a dysregulated increase in approach motivation, and depression (i.e., extreme negative affect) with increased behavioral inhibition and avoidance (Kasch et al., 2002; Meyer et al., 2001, 2007; Trew, 2011; Urošević et al., 2008). We therefore also conducted several exploratory analyses investigating how any effects of positive/negative affect on Pavlovian-instrumental interactions may be moderated by individual-difference factors related to approach and avoidance motivation, including depression/hypomania, behavioral activation/inhibition (Corr et al., 1995), and gender (De Carli et al., 2017; Robinson & Sahakian, 2009).

Experiment 1

Method

Participants

Participants were 122 adults (53 women, 67 men, 2 who did not endorse a binary gender[1]; mean age 39.01 years, range 22-70) from the United States and Canada, recruited online on Amazon Mechanical Turk (MTurk) via the CloudResearch interface. This study was approved by the Institutional Review Board of Princeton University, and all participants provided written informed consent. Total study duration was approximately 35 minutes, and participants received a base payment plus a bonus proportional to task winnings to ensure that their choices in the task were incentive-compatible (mean total payment = $6.65, SD = 1.03). Given the exploratory nature of Experiment 1, sample size was determined based on a survey of related research assessing individual differences in the reward/punishment Go/No-Go task (Csifcsák et al., 2020; Mkrtchian et al., 2017; Swart et al., 2018).

Procedure

Participants completed a behavioral task designed to measure the effects of a video-based affect induction procedure on Pavlovian-instrumental interactions in learning. This was followed by a short demographic survey and two self-report surveys: the 7-Up 7-Down scale (a 14-item measure of trait hypomania and depression; Youngstrom et al., 2013), and the shortened Behavioral Inhibition System/Behavioral Approach System scale (BIS/BAS; a 12-item measure of trait behavioral inhibition and activation; Carver & White, 1994; Pagliaccio et al., 2016). All tasks and surveys were presented using the jsPsych library for JavaScript (De Leeuw, 2015),

[1] Following a convention set in part by reporting requirements by the National Institute of Mental Health, we asked participants to report their gender with response options of 'male' or 'female'. We acknowledge that this is a misuse of terms given that male/female are sex terms, not gender terms. We discuss participants' self-report as "gender" and discuss "gender difference" in spite of the erroneous use of sex terms in our demographic form, because we reason that participants were likely to interpret this question as asking them to report their self-identified gender, not their biological sex (which we did not assess).

along with custom-written server code (available at https://github.com/nivlab/nivturk) using the Flask software package for Python.

Reward/punishment Go/No-Go task. In the behavioral task (a modified version of a task originally developed by Guitart-Masip et al., 2011; see Figure 1A), participants were repeatedly shown different 'robot' stimuli. Robots were depicted as travelling down a conveyer belt into a 'scanner'; during the 1.5 second response window while a robot was in the scanner the participant could either press the space bar (active 'Go' response), or not press a key (inhibitory 'No-Go' response). Participants were informed that they would observe different robot types, and that their task was to learn the correct response (Go vs. No-Go) for each robot type based on feedback (points won/lost) following each action. The instructions presented to participants can be found in Supplementary Section S1.

There were four 'types' of robot stimuli, visually denoted by different rune images on robots' breastplates (see Figure 1; mapping between rune image and stimulus type was pseudo-randomized across participants). The four stimulus types differed in terms of their payout domain (gain versus loss; explicitly signaled to the participant by a blue or yellow 'scanner light' that appeared at the same time as the robot) and their correct action (Go versus No-Go; unsignalled, and learned from trial and error), as depicted schematically in Figure 1B. Gain-domain stimuli provided payouts of either +10 or +1 points, whereas loss-domain stimuli provided payouts of either -1 or -10 points. For each stimulus, participants probabilistically received the better of the two possible payouts if they made the correct action (80% chance of better payout, 20% chance of worse payout), and the worse of the two possible payouts if the chosen action was incorrect (80% chance of worse payout, 20% chance of better payout). As in earlier studies using the reward/punishment Go/No-Go task (e.g., Cavanagh et al., 2013; Geurts et al., 2013; Guitart-

Masip et al., 2011, 2012; Swart et al., 2017), this orthogonalized design resulted in four robot types (see Figure 1C): gain-domain robots for which the correct response was 'Go' (henceforth, Go to Win robots [GW]), gain-domain robots for which the correct response was 'No-Go' (No-Go to Win [NGW]), and equivalent robots in the loss domain (Go to Avoid Losing [GAL] and No-Go to Avoid Losing [NGAL]). The scanner light color, which was directly associated with winning or losing points regardless of the robot or action chosen, is therefore a Pavlovian cue that is associated with either the appetitive or the aversive domain of the task. The specific color (yellow or blue) associated with each domain was randomly determined for each participant; the relationship between color and domain was explicitly instructed and stayed constant throughout the task. Of the four stimulus types, GW and NGAL stimuli were instrumental-Pavlovian 'congruent' since there was a match between the correct response and Pavlovian approach/avoidance tendencies for each, whereas NGW and GAL stimuli were 'incongruent' (mismatch between Pavlovian and instrumental response tendencies).

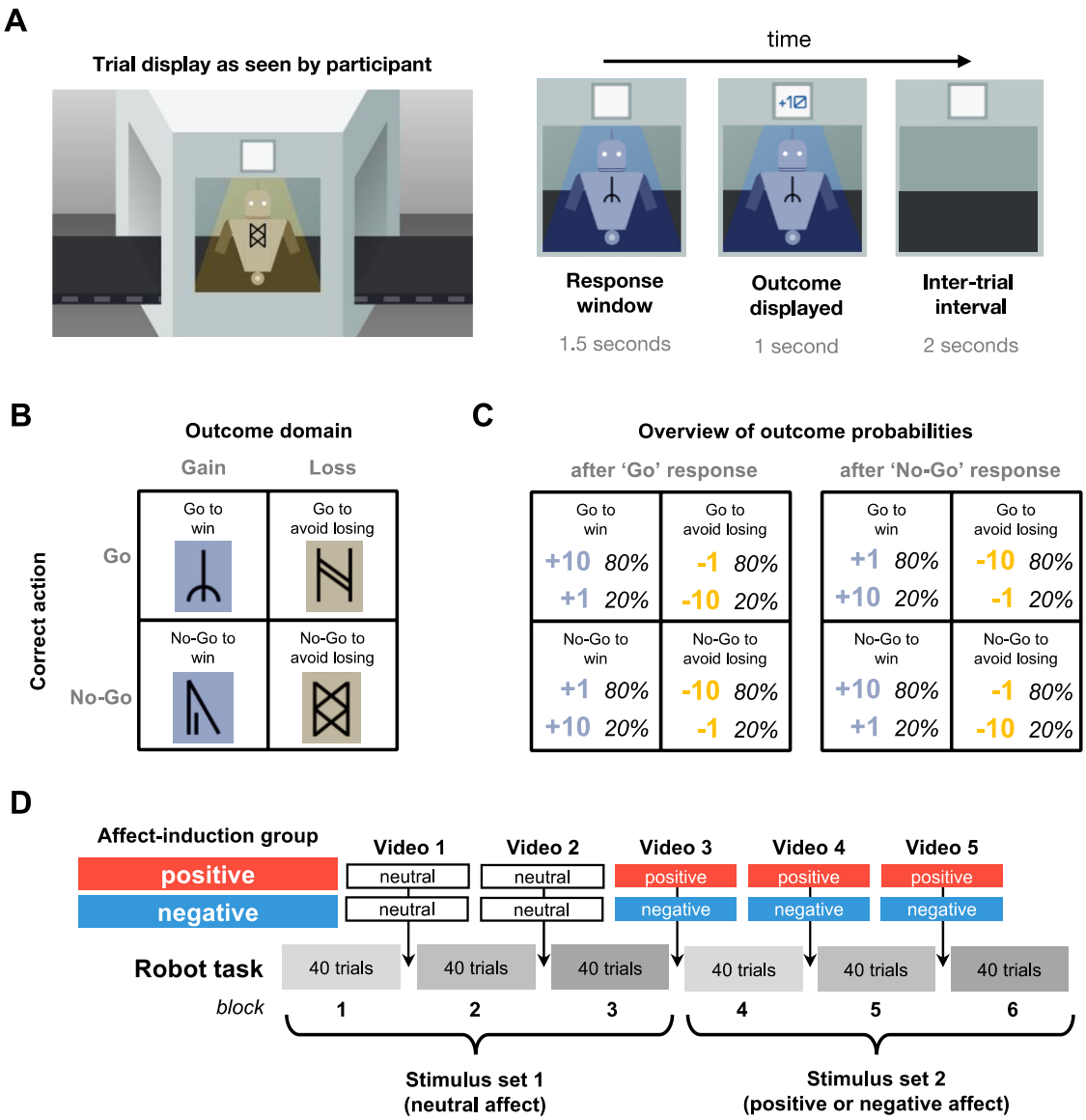


Figure 1. (A) **Trial schematic**. On each trial, a robot entered the 'scanner' from the left of screen, prompting a response (Go or No-Go) from the participant during a 1.5 second response window. The outcome (number of points won) was subsequently presented for 1 second, followed by an inter-trial interval animation in which the conveyer belt carried the old robot out of view and a new robot into the scanner. *(B)* **The four stimulus types**, produced by a factorial combination of outcome domain (Gain/Loss) and action (Go/No-Go), each indicated by a unique rune symbol. The color of the scanner light denotes outcome domain (randomized across participants; for illustration purposes blue denotes gain and yellow denotes loss). *(C)* **Outcome probabilities** for each stimulus type after a Go response (left) or after a No-Go response (right), depending on stimulus type. Each stimulus was 80% predictive of the better possible outcome (within the outcome domain) given a correct choice. *(D)* **Overall procedure**. The task comprised 6 blocks of 40 trials each. Between blocks, participants watched 90-second film clips with neutral emotional content (stimulus set 1) or either positive or negative emotional content (stimulus set 2). For both affect-induction groups, the two videos that played during the first stimulus set were neutral. The three videos played during the second stimulus set were either all positive or all negative, depending on the participant's assigned condition.

In line with recent studies using this task (but contrary to the original task by Guitart-Masip et al., 2011), the payout domain (gain/loss) of a robot was explicitly indicated to the participant on each trial by either a blue or yellow scanner light that illuminated the stimulus during the response window (Swart et al., 2017, 2018). Participants received prior instruction and training in this color/domain mapping (which was pseudorandomized across participants), and were also instructed that feedback was probabilistic. Participants were shown eight stimuli in total across the task, divided into two sets of four stimuli each (with one robot of each type within each stimulus set). The first stimulus set was presented in task blocks 1-3, and the second stimulus set in task blocks 4-6 (i.e., participants were required to learn a new set of rune-action associations at the start of block 4). Participants were informed at the start of the task that robots would be distinguished by different rune symbols, and were not provided with any additional instructions between stimulus set 1 and stimulus set 2 (i.e., between block 3 and block 4).

Affect induction procedure. We used a video-based affect-induction procedure to induce either positive affect or negative affect (between-participants design) during the second stimulus set. Video-based affect inductions have been shown to induce robust changes in participants' self-reported mood in both in-person and online experiments (Ferrer et al., 2015; Joseph et al., 2020). We used a set of affect-induction film clips that we have validated in our previous work (Bennett, Radulescu, et al., 2021).

Based on the literature, we expected to observe considerable inter-individual differences in the strength of Pavlovian influence on participants' instrumental behavior (Csifcsák et al., 2020; Mkrtchian et al., 2017; Swart et al., 2018). To control for these individual differences when assessing the effect of the affect induction, we used a change-from-baseline procedure (see Figure 1D) in which participants completed the first stimulus set at their baseline affect level

(i.e., without an active affect induction), and the second stimulus set in either an induced positive or an induced negative affective state. Specifically, each of the three blocks in the second stimulus set was preceded by a distinct affect-induction video clip (i.e., either three distinct happy video clips or three distinct sad video clips, depending on which condition a participant was in). To control for any non-specific effects that watching video clips would have on behavior unrelated to the clips' emotional content (e.g., distraction due to interruption of ongoing task performance), the first and second blocks of the first stimulus set were followed by distinct *neutral* video clips to ensure that task demands were matched as closely as possible between the two stimulus sets. To verify the efficacy of the affect induction, we collected participants' self-reports of their current mood (both valence and arousal) before and after each video using an Affective Slider (Betella & Verschure, 2016).

Data analysis

Choice behavior was analyzed using generalized linear mixed-effects analyses (logistic link function), with response accuracy as the dependent variable (correct coded as 1, incorrect coded as 0). Within these models, we assessed the effects of different stimulus properties (e.g., payout domain, correct action type) on accuracy, as well as whether the affect induction had either a main effect on accuracy or interacted with particular stimulus properties (e.g., facilitating learning from 'Go' outcomes specifically). Self-reported mood ratings were analyzed separately using linear mixed-effects analyses. Analyses were conducted using the *lme4* package (Bates et al., 2015) and the *lmerTest* package (Kuznetsova et al., 2017) in R. All mixed-effects models incorporated random intercepts for participants as well as random slopes for all main effects and interactions that were entirely within-participant (Barr et al., 2013). *p*-values were calculated using the Satterthwaite degrees-of-freedom approximation for individual coefficients within

linear mixed-effects analyses, and the Wald *t*-to-*z* test for omnibus linear tests and logistic mixed-effects analyses (Meteyard & Davies, 2020). Finally, we also tested whether any self-report survey measures moderated the effect of the affect induction on learning by assessing the correlations between participant-wise random slopes for each effect and interaction from fitted mixed-effects models and sum scores from self-report measures (using Spearman rank-order correlations with a false-discovery-rate correction for multiple comparisons).

Since we had no specific hypotheses about the effect of the affect induction on behavior in the reward/punishment Go/No-Go task prior to analysis of Experiment 1, all Experiment 1 analyses were deemed exploratory, and findings were treated as preliminary pending replication in Experiment 2. In the absence of specific hypotheses, the sample size for Experiment 1 was determined based on effect size estimates derived from recent studies using similar variants of this task (Mkrtchian et al., 2017; Moutoussis et al., 2018; Swart et al., 2017). Full correlation and mixed-effects regression tables, including details of predictor and outcome coding schemes, can be found in the Supplementary Material, Section S2; see Experiment 2 methods for its power analysis. All statistical assumptions were met for each of the analyses reported in this manuscript.

Data quality control

Because of concern about the potential for data contamination due to careless responding from online participants, we excluded data from all participants whose data failed quality-control tests. For questionnaire data, we measured per-questionnaire mean response times, and excluded participants who responded excessively rapidly (< 1 second per question for the BIS/BAS or < 3 seconds per question for the 7-Up 7-Down; different criteria were used for each scale because of their different average question lengths; Ophir et al., 2020). For the behavioral task, since

Pavlovian-instrumental interactions are dependent on participants' ability to discriminate stimuli from different payout domains (i.e., blue versus yellow scanner lights), participants who self-reported color blindness were also excluded from analysis. We also excluded participants who did not show above-chance task learning (according to a one-tailed binomial test against chance, $\alpha = .05$) for GW stimuli (the easiest condition to learn; Albrecht et al., 2016; Guitart-Masip et al., 2012). To ensure that participants had not disengaged entirely from the task, we also excluded those who made more than 75% 'No-Go' responses across the task. Finally, to ensure that participants engaged adequately with each affect-induction video clip, we embedded several attention checks within the delivery of each video: after each video, participants were required both to answer a simple comprehension check question about the content of the video and to identify a still frame from the video from among an array of foils. Participants who responded incorrectly to more than one of these comprehension checks across the entire task were excluded from further analysis. We also recorded participants' interactions with their web browser and excluded from analysis any participant who clicked away from the videos for more than 10 seconds per video on average. As a result of these exclusion criteria, 31 participants (25.4% of sample) were excluded by an analyst blind to participant condition, after which 91 participants remained for analysis.

Results

The affect-induction procedure was successful at altering self-reported mood

We first verified that the affect-induction procedure successfully induced changes in self-reported mood. As expected, we found a significant effect of video condition on change in mood valence in blocks 4-6, χ^2 (2) = 120.60, $p < .001$ (post-video minus pre-video self-report; see Figure 2A). This effect was driven both by mood improvement after happy videos relative to neutral videos ($\beta = 0.10$, $p < .05$), and by mood deterioration after sad videos relative to neutral videos ($\beta = -0.26$, $p < .001$). There was no significant main effect of block, and no significant interaction between video condition and block.

We also found a significant effect of the affect induction on self-reported mood arousal, $\chi^2(2) = 6.34$, $p < .05$, driven by a tendency for happy videos to increase arousal more than neutral videos, ($\beta = 0.09$, $p < .05$; see Supplementary Figure S1). There was no difference between arousal change following neutral videos and arousal change following sad videos ($\beta = 0.03$, $p = .50$) and, as with valence, there was no significant effect of block number on change in arousal, and no significant interaction between video condition and block number.

Pavlovian-instrumental interactions were evident in the data

Next, we verified the standard pattern of Pavlovian-instrumental interactions in our data. The typical manifestation of this effect is that participants show better performance for congruent stimuli (GW and NGAL) than incongruent stimuli (GAL and NGW). In line with this finding, we found a significant interaction between outcome valence and correct action (χ^2 (1) = 80.55, $p < .001$), driven by increased accuracy for congruent stimulus types (GW and NGAL) relative to incongruent stimulus types (GAL and NGW; see Figures 2B, C). There was also a significant

effect of block number on choice accuracy (χ^2 (1) = 114.59, p < .001), with accuracy improving over time within each stimulus set (β = 0.80, p < .001).

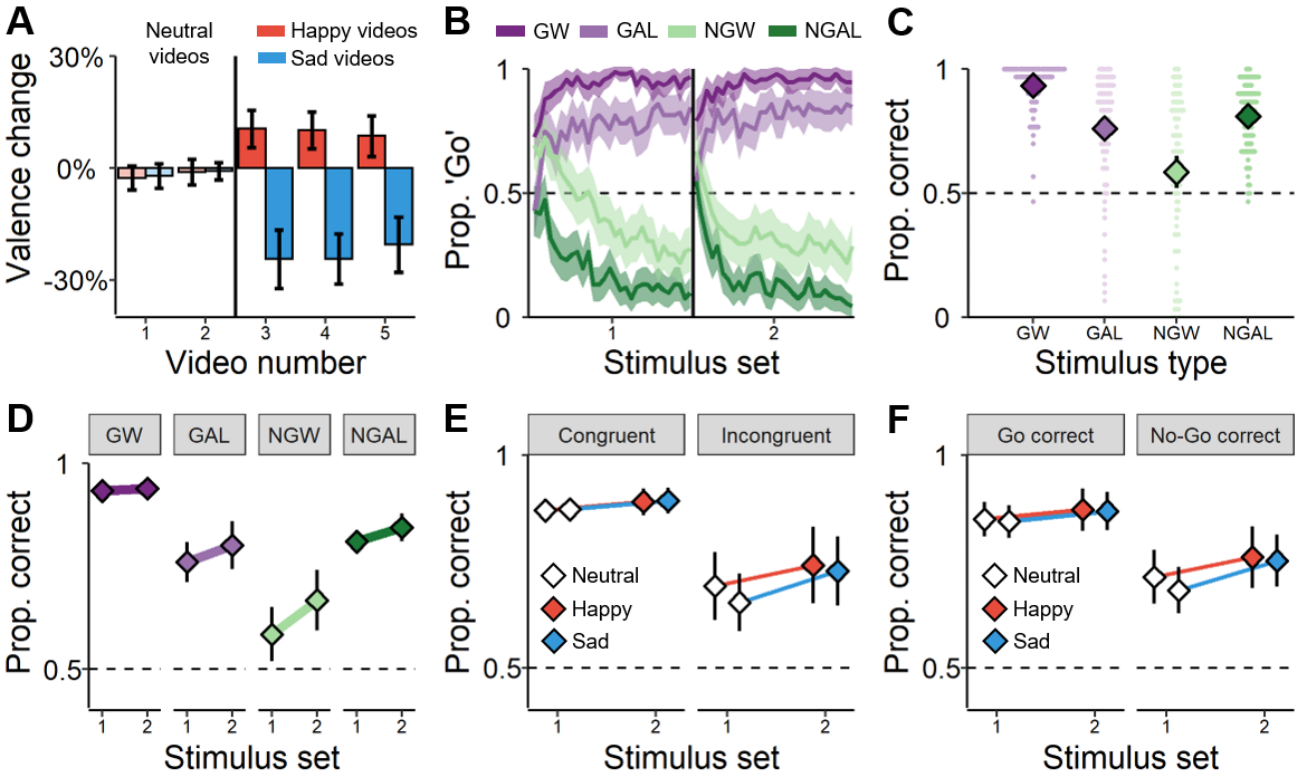


Figure 2. **Overview of results for Experiment 1**. *(A)* **Mean change in mood valence produced by each video show the affect manipulation was effective** (post-video minus pre-video valence as a percentage of the total length of the Affective Slider). The solid vertical line demarcates the boundary between stimulus set 1 (left, in which only neutral videos were presented) and set two (right, in which participants viewed either three happy or three sad videos). *(B)* **Mean proportion of 'go' responses over time for each stimulus type demonstrate a Pavlovian-instrumental interaction** (GW: go-to-win, GAL: go-to-avoid-losing, NGW: no-go-to-win, NGAL: no-go-to-avoid-losing). Darker colors denote congruent Pavlovian-instrumental stimuli, and lighter colors denote incongruent stimuli. *(C)* **Overall proportion correct for each stimulus type in stimulus set 1** (i.e., neutral affect condition). Diamond markers denote overall means, and background points denote condition means for individual participants. *(D)* **Change in proportion correct between stimulus set 1 and 2**, averaged across all participants. Performance improvement between stimulus set 1 and 2 was greater for stimulus types in which mean performance was worse in set 1. *(E,F)* **Change in proportion correct due to the affect manipulation:** Proportion correct as a function of *(E)* stimulus-response congruency (congruent: GW and NGAL; incongruent: GAL and NGW) or *(F)* correct response type (Go: GW and GAL; No-Go: NGW and NGAL) and stimulus set, presented separately for positive and negative affect conditions. Error bars/shading denote the 95% confidence interval of the mean. Note that some error bars are small enough to be obscured by condition markers.

In addition, we found a significant main effect of stimulus set (χ^2 (1) = 17.97, p < .001), driven by overall improved accuracy in the second stimulus set and consistent with a generalized practice effect (independent of any effects of the affect induction). We also found evidence for a significant three-way interaction between stimulus set, outcome valence, and correct action, such that the strength of the Pavlovian-instrumental interaction effect *decreased* from the first to the second stimulus set, (χ^2 (1) = 4.79, p < .05). As shown in Figure 2D, this effect was driven by greater performance improvement in the second stimulus set for the Pavlovian-incongruent NGW and GAL stimuli (which were further from ceiling performance in the first stimulus set) than for the Pavlovian-congruent GW and NGAL stimuli (post-hoc test: χ^2 (1) = 7.22, β = 0.47, p < .01).

No overall effects of affect induction on Go/No-Go choices

To test how the affect induction influenced task behavior, we repeated the mixed-effects analysis described above while including additional coefficients corresponding to (a) a main effect of affect induction condition (positive versus negative) as well as (b) all two-, three- and four-way interactions between affect condition and other predictors (e.g., payout domain, correct action). In each case, we found no evidence for an effect of the affect induction on behavior (all p > .12). This indicated that, contrary to our expectations, there was no evidence that participants who received a positive affect induction in the second stimulus set showed any systematic differences in task behavior relative to participants who received a negative affect induction. This lack of effect of the affect induction encompassed several aspects of behavior that we had hypothesized might be influenced by an affect induction, including the strength of the Pavlovian-instrumental interaction (Figure 2E) and overall go versus no-go response prepotency (Figure 2F).

Finally, we considered the possibility that, although there were no significant group-level main effects or interactions associated with the affect induction, individual differences in depression, mania, behavioral inhibition, or behavioral activation might nevertheless correlate with the size of the affect inductions' effect on behavior. However, we found no significant association between any individual difference measure and participant-wise random slopes for any effects or interactions, for either positive or negative affect inductions (all $p > .09$, FDR-corrected; see Supplementary Materials for full correlation tables).

Interim Discussion

Although manipulation-check analyses revealed that the positive and negative affect inductions had the expected effect on participants' self-reported mood in both conditions, we found no evidence that Go/No-Go behavior differed between participants who received a positive affect induction and those who received a negative affect induction. This was unexpected given previously observed effects of affective states on prepotent action tendencies and perception of emotionally-relevant stimuli (e.g., Bouhuys et al., 1995; Lang et al., 1990). We therefore reasoned that our study might not have had sufficient statistical power to detect a small effect of affect condition on approach/avoidance behavior or Pavlovian-instrumental interactions. To investigate this possibility, we next conducted a preregistered replication of Experiment 1 in a larger sample of participants with increased statistical power for detecting small effects (see Experiment 2 Method section for power analysis).

Experiment 2

Experiment 2 was a preregistered replication of Experiment 1 using an identical behavioral task design. However, between Experiment 1 and Experiment 2 we made several

modifications to our online participant recruitment protocol based on separate quality control assays conducted after data collection for Experiment 1 had been completed (see Zorowitz et al., 2021). Briefly, this assay revealed (a) that for the specific recruitment procedure that we were using in our laboratory at this time, average data quality for participants recruited through Prolific was better than for participants recruited via MTurk, and (b) that improved questionnaire screening aided in the detection of low-effort participant responding. For these reasons, Experiment 2 recruited participants via Prolific rather than MTurk and used a two-stage recruitment process in which participants first completed a questionnaire battery, inviting back only participants who passed a series of attention checks to complete the reward/punishment Go/No-Go task (see Supplementary Information for full details of recruitment procedure). In total, 25.2% of an initial screening sample did not give correct responses to the attention checks embedded in the questionnaires, and were therefore excluded from re-recruitment to complete the behavioral task.

Methods and analyses for Experiment 2 were pre-registered prior to data collection. Our pre-registered hypotheses were that we would replicate the results of Experiment 1 with respect to the overall effects of the affect induction on mood, the overall Pavlovian-instrumental interaction, and the null effects of the affect induction on Pavlovian-instrumental interactions (preregistration document available in project OSF repository). Full correlation and mixed-effects regression tables, including details of predictor and outcome coding schemes, can be found in the Supplementary Material, Section S5.

Method

Participants

Prior to data collection, we determined a target sample size (after exclusions) of 300 participants for Experiment 2 by simulating from a mixed-effects model fit to data from Experiment 1 using the *simr* package in R (Green & MacLeod, 2016). Specifically, we estimated the sample size necessary to have 80% power to detect (α = .05) an effect size of the affect induction on choice behavior of β = 0.5 (corresponding to a 3% difference in the effects of the positive and negative affect inductions on proportion correct for any stimulus type). This compares to a post-hoc power estimate for Experiment 1 of approximately 43% for this effect size. See online project repository (https://osf.io/zm57r/) for further details of power analysis.

As a result of lower-than-estimated rates of participant exclusion after applying the same set of exclusion criteria as in Experiment 1 (exclusion rate of 16.3% compared to 25.4% in Experiment 1), we marginally exceeded our target sample size of 300, and retained a final sample of 335 participants for analysis (181 women, 150 men, 4 who did not endorse a binary gender; mean age 33.85 years, range 18-74). As in Experiment 1, participants who completed the behavioral task received a task payment that included an incentive-compatible bonus for task performance (mean total payment = USD $5.49, SD = 0.31).

Materials

All task and questionnaire materials were identical between Experiment 1 and Experiment 2, with the exception that participants also completed additional self-report measures of anxiety (the GAD-7; Spitzer et al., 2006), worry (the 3-item abbreviated Penn State Worry Questionnaire (PSWQ); Kertz et al., 2014), and anhedonia (the Snaith-Hamilton Pleasure Scale (SHAPS); Snaith et al., 1995). Since we did not collect data on these measures in Experiment 1, analyses of correlations between these scales and task behavior in Experiment 2 were treated as exploratory.

Results

All results reported below (with the exception of those in the *Exploratory Analyses* subsection) were preregistered confirmatory tests. In each section, we first report all results that were consistent with results from Experiment 1, before reporting any discrepancies between experiments. As with Experiment 1, full regression and correlation tables are available in the Supplementary Material.

The affect-induction procedure was successful at altering self-reported mood

As in Experiment 1, the affect induction successfully modulated the valence of participants' self-reported mood, χ^2 (2) = 308.25, $p < .001$ (Figure 4A). Follow-up post hoc tests confirmed that this effect was driven both by mood improvement after happy videos ($\beta = 0.13$, $p < .001$), and mood deterioration after sad videos ($\beta = -0.18$, $p < .001$). There was also a significant effect of video condition on change in self-reported mood arousal (χ^2 (2) = 23.71, $p < .001$) driven by increases in arousal following happy videos condition relative to neutral videos ($\beta = 0.08$, $p < .001$; see Supplementary Figure S1).

Unlike in Experiment 1, in Experiment 2 we found a significant interaction between video condition and block number (χ^2 (2) = 6.14, $p < .05$), which post-hoc tests revealed was driven by a reduction in the effect of the negative affect induction over successive blocks ($\beta = 0.04$, $p < .05$).

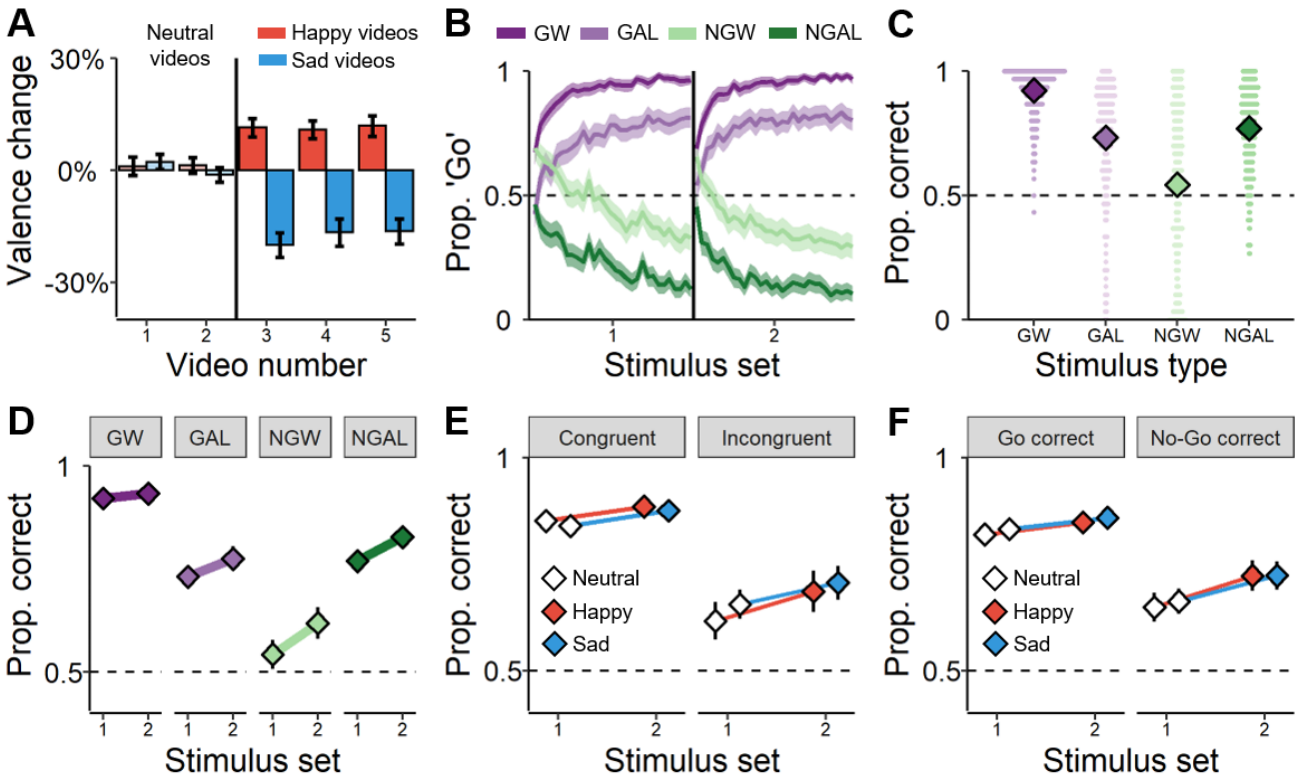


Figure 3. **Overview of results for Experiment 2.** *(A)* **Mean change in mood valence produced by each video show the affect manipulation was effective** (post-video minus pre-video valence as a percentage of the total length of the Affective Slider). The solid vertical line demarcates the boundary between stimulus set 1 (left, in which only neutral videos were presented) and set two (right, in which participants viewed either three happy or three sad videos). *(B)* **Mean proportion of 'go' responses over time for each stimulus type demonstrate a Pavlovian-instrumental interaction** (GW: go-to-win, GAL: go-to-avoid-losing, NGW: no-go-to-win, NGAL: no-go-to-avoid-losing). Darker colors denote congruent Pavlovian-instrumental responses, and lighter colors denote incongruent responses. *(C)* **Overall proportion correct for each stimulus type in stimulus set 1** (i.e., neutral affect condition). Diamond markers denote overall means, and background points denote condition means for individual participants. *(D)* **Change in proportion correct between stimulus set 1 and 2,** averaged across all participants. *(E,F)* **Change in proportion correct due to the affect manipulation**: Proportion correct as a function of *(E)* stimulus-response congruency (congruent: GW and NGAL; incongruent: GAL and NGW) or *(F)* correct response type (Go: GW and GAL; No-Go: NGW and NGAL) and stimulus set, presented separately for positive and negative affect conditions. Error bars/shading denote the 95% confidence interval of the mean. Note that some very narrow error bars are obscured by condition markers.

Pavlovian-instrumental interactions were evident in the data

We once again found evidence for a Pavlovian-instrumental interaction in learning (χ^2 (1) = 222.37, $p < .001$), with participants showing decreased accuracy for Pavlovian-incongruent

stimuli relative to Pavlovian-congruent stimuli (Figures 4B, C). Here, too, general accuracy significantly improved from the first to the second stimulus set (χ^2 (1) = 25.83, p < .001). Unlike in Experiment 1, however, we found no evidence that the strength of Pavlovian-instrumental interactions differed between the first and the second stimulus set (χ^2 (1) = 0.73, p = .39).

No overall effects of affect induction on Go/No-Go choices

As in Experiment 1, even with the increased sample size there were no statistically significant main effects or interactions involving the affect induction, (all p > .52; see Figure 4E and 4F). Unlike in Experiment 1, however, in Experiment 2 we found a significant negative correlation between self-reported behavioral inhibition (BIS subscale) and participant-wise random slopes for the effect of the positive affect induction on gain-domain stimuli (Spearman ρ = -.23, p = .02, false-discovery-rate corrected for multiple comparisons; see Supplementary Figure S2). The negative sign of this correlation indicates that, after a positive affect induction, participants with higher BIS scores showed a smaller improvement on gain-domain stimuli (i.e., GW and NGW), relative to participants with lower BIS scores. No other correlations between survey scores and affect-related random slopes were statistically significant (see Supplementary Material).

Exploratory analyses reveal age and gender effects

As an additional control analysis, we tested whether differences in the randomized light color associated with the gain vs. the loss domain (blue versus yellow) were associated with differences in task behavior. Using an additional mixed-effect logistic regression analysis, we found no evidence for an effect of this factor on task behavior (see Supplementary Table S11). In addition, we sought to account for the possibility that the effects of affect inductions on mood may have been relatively short-lived, and that our whole-stimulus-set analysis might have

obscured an effect of the affect induction on behavior in trials immediately following the affect induction. To address this possibility, we repeated all analyses while excluding trials from the second half of each block in set 2 (i.e., excluding trials that took place after affect inductions might plausibly have 'worn off'). As in the main analyses reported above, however, there were no main effects or interactions of affect induction on task behavior in this more restricted analysis.

Finally, we conducted several exploratory analyses to test the association between demographic factors and the effects of the affect induction on task behavior. First, using the same random-effects-correlation analysis described above for self-report scales, we found evidence for a small positive association (Spearman $\rho = .19$, $p < .05$, false-discovery-rate corrected) between age and the effect of the positive affect induction on performance for gain-domain stimuli, such that older participants tended to show a greater improvement in performance on gain-domain stimuli after a positive affect induction than younger participants (see Supplementary Figure S2). In addition, inclusion of binary self-reported gender (man/woman) as a factor in mixed-effects analysis revealed that the size of the affect-related change in the strength of Pavlovian-instrumental interactions differed between men and women participants ($\chi^2(1) = 7.12$, $p < .01$). Given the relatively complex patterns of differences between genders in the effects of the affect induction (see Supplementary Figure S3), we turned to computational modelling of behavior to explicate this result.

Interim Discussion

Broadly, Experiment 2 replicated the null findings of Experiment 1: although performance on the task changed (improved) from the first to the second stimulus set, there was no evidence

for significant differences in the pattern of these changes between participants who received a positive affect induction and those who received a negative affect induction. In our exploratory analyses, however, we found some evidence that the strength of the effects of the affect inductions on behavior may have been moderated by participants' demographic characteristics (specifically, age, gender, and self-reported behavioral inhibition).

Computational modelling

We used computational modelling to investigate two aspects of the results that remained unresolved given the analyses reported in previous sections. First, although standard inferential statistics may fail to reject a null hypothesis, formal comparison of computational models can better quantify the strength of evidence in favor of the null hypothesis versus alternative hypotheses. An alternative approach to this question would be to estimate Bayes Factors, which directly compute the strength of evidence for the null hypothesis versus the alternative hypothesis in a Bayesian framework (Kass & Raftery, 1995). However, there is currently no consensus as to the best approach for computing Bayes Factors for mixed-effects regression analyses like those reported above (see van Doorn et al., 2021). We therefore elected to use trial-by-trial computational modelling of data to compute the strength of evidence for the null hypothesis that affect inductions did not substantially influence task behavior. Like the Bayes Factor approach, this model comparison approach allowed us to compare the strength of evidence for different models (including a null model). We compared models using the WAIC statistic, which approximates the estimated out-of-sample-prediction error (Gelman et al., 2014). Second, we used participant-level parameter estimates to gain a better understanding of moderating effects that were suggested by exploratory analyses of Experiment 2. Full details

regarding the computational modelling methodology that we employed can be found in the Supplementary Material (Section S6).

Corroborating the model-agnostic results reported in previous sections, the best-fitting model was Model 2 (see Table 1). In Model 2, parameters were free to change between the first and second stimulus set (e.g., as a result of practice effects), but without any group-level differences between the positive and negate-ve affect inductions. Model 3 provided a statistically equivalent fit to the data (i.e., the difference in the WAIC values for Models 2 and 3 was less than the standard error of the estimated difference in WAICs), but was less parsimonious than Model 2; this indicates that the additional parameters in Model 3 allowing for group-level differences between the positive and negative affect conditions did not account for a meaningful amount of variance in the data. Our data therefore suggest that accounting for the valence of the affect induction (positive or negative) did not add any additional explanatory power over a simpler model in which parameters were simply allowed to vary as a function of time on task.

Table 1. Summary of model comparison

Model	*N* parameters estimated (total)	WAIC	ΔWAIC (SE)
1	1,712	68,683.8	3,955.1 (132.3)
2	2,996	64,728.7	0 (0)
3	3,002	64,729.1	0.4 (4.3)

Note: WAIC values are presented on a deviance scale (lower numbers indicate better model fit).

Interpretation of model parameters

Ultimately, for the best-fitting model (Model 2) we found that the only parameter for which there was a credible difference in parameter estimates between stimulus set 1 and 2 was the learning rate parameter η. As seen in Figure 4, on average learning rates were higher in

stimulus set 2 than in set 1, consistent with the practice effects seen in Figures 2D and 3D. There were no credible between-set differences in either the go-bias parameter *b* or, importantly, in the Pavlovian-bias parameter π.

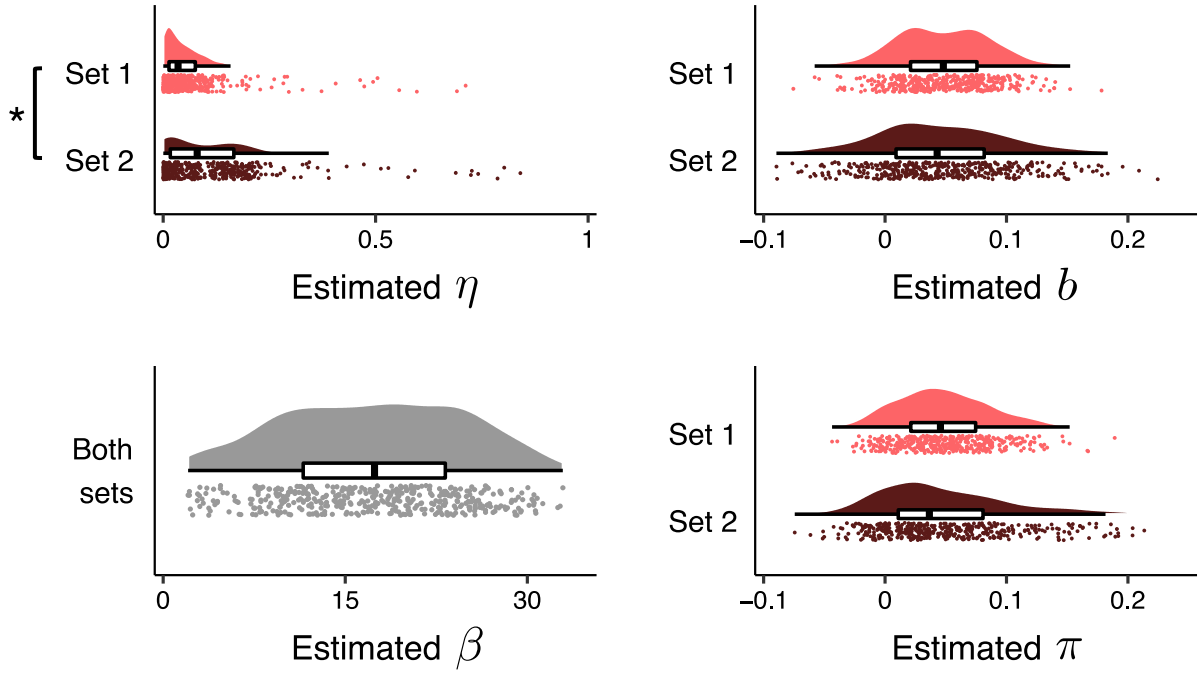


Figure 4. **Estimated parameters from the best-fitting computational model** (Model 2). Scatterplots depict point estimates (posterior medians) of each parameter for each participant separately for stimulus set 1 (pink) and stimulus set 2 (brown). The β parameter (grey) was constrained to be equal across both stimulus sets. Box-and-whisker plots depict the median and interquartile range of participant-wise parameter estimates. * denotes a credible between-set mean difference in estimated parameter values (95% Bayesian HDI of difference excludes zero).

We next investigated between-gender parameter differences in the first stimulus set and found that, compared to men, women tended to show a credibly stronger go bias (mean baseline *b* for women: 0.054; for men: 0.041; Cohen's *d* = 0.34) as well as a stronger Pavlovian bias (mean baseline π for women: 0.051; for men: 0.045; Cohen's *d* = 0.17). There were no credible baseline differences between genders in either the learning-rate parameter η or the softmax inverse temperature β. See supplementary information for additional correlation matrices

between self-report survey scores (Supplementary Table S12) and between model parameter estimates (Supplementary Table S13).

Self-reported behavioral inhibition and gender moderate affect-induction effects

Next, we investigated whether any participant-level covariates moderated the strength of the effect of the affect induction on any parameter (within Model 2; see Table 1). In this analysis, we found credible evidence (99.5% HDIs excluding zero) for two small moderating effects involving self-reported behavioral inhibition (BIS subscale of BIS/BAS). Specifically, self-reported BIS moderated the effects of the affect induction on both the go-bias parameter b and the Pavlovian bias parameter π (i.e., there was a credible difference in the correlation between BIS and the changes in b and π after a positive affect induction and the correlation between BIS and changes in these parameters after a negative affect induction). Participants reporting higher trait behavioral inhibition scores tended to show a greater reduction in the strength of the go bias and the Pavlovian bias after a positive affect induction (correlation with change in go bias: Spearman ρ = -.07, 95% HDI [-.17, .02]; with change in Pavlovian bias: ρ = -.16, 95% HDI [-.26, -.06]) and an increase in the strength of these parameters after a negative affect induction (go bias: ρ = .11, 95% HDI [.02, .20]; Pavlovian bias: ρ = .10, 95% HDI [.01, .19]).

Finally, to explicate the complex gender interactions that were observed in the mixed-effects analyses, we investigated whether gender moderated any of the effects of the affect induction on model parameters. In line with the significant role of gender as a moderator in Experiment 2, we found there to be a credible difference between genders in the effect of the affect induction on the Pavlovian bias. As shown in Figure 5, this moderating effect was such that women tended to show a stronger Pavlovian bias (i.e., increased π) after the negative affect

induction, whereas men tended to show a weaker Pavlovian bias (i.e., decreased π) after the negative affect induction. For both genders there was no credible effect of the positive affect induction on Pavlovian bias, and there were no credible moderating effects of gender for any other model parameter.

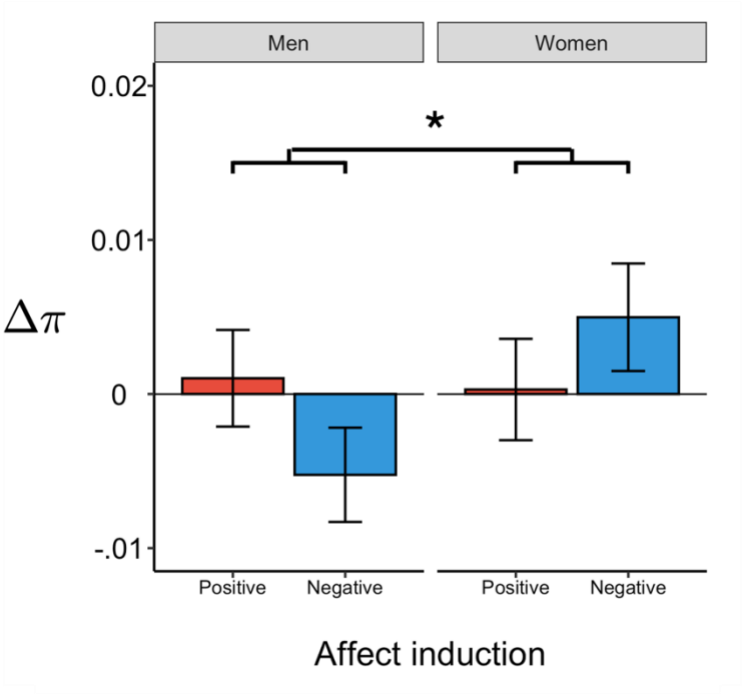


Figure 5. Estimated affect-related change in Pavlovian bias parameter ($\Delta\pi$) as a function of affect-induction condition (positive affect: red; negative affect: blue) and participant gender. Error bars denote the standard error of the mean. * denotes a credible interaction between gender and affect-induction condition (i.e., 95% Bayesian HDI for interaction effect excludes zero).

General Discussion

In this study, we used a video-based affect-induction procedure to test whether induced positive and negative affect modulated Pavlovian influences on instrumental learning of approach/avoidance behavior. Across two experiments – an exploratory Experiment 1 and a larger preregistered replication in Experiment 2 – we found no evidence for any main effects of induced positive or negative affect on either overall approach/avoidance tendencies or

Pavlovian-instrumental interactions. This result was corroborated by formal comparison of computational cognitive models: in the model that fit participants' behavior best, model parameters changed over time (accounting for non-specific practice effects), but there were no differences in patterns of parameter change between participants who received a positive affect induction and those who received a negative affect induction.

The lack of an effect of induced positive and negative affect on Pavlovian-instrumental interactions is surprising in the context of influential theories of emotion that link positive and negative affect with approach and avoidance motivation respectively (Cacioppo et al., 1999; Lang, 1995; Watson et al., 1999). We could have expected either an overall enhancement of approach (withdrawal) for positive (negative) affect, or a potentiating effect on the Pavlovian influences in the relevant affective domain (e.g., positive affect enhancing the effect of appetitive Pavlovian stimuli on instrumental responding). However, we did not find evidence for either of these possibilities. As such, our null effect is conceptually inconsistent with a body of work showing affective modulation of defensive startle reflexes (e.g., Bradley et al., 1990; Lang, 1995; Lang et al., 1990; Vrana et al., 1988), though we emphasize that reflex behaviors are behaviorally and neurophysiologically distinct from the volitional instrumental behavior that our task measured (Balleine, 2019). Our results are in line, however, with a previous study using an approach/avoidance task, which found that induced positive/negative affect did not influence approach or avoidance of emotional face stimuli (Vrijsen et al., 2013). More broadly, future research in this area could consider increasing the psychological salience of approach behaviors by requiring participants to touch cues presented on a touchscreen, rather than pressing a spatially non-contiguous space bar as in the present study.

More broadly, our findings resonate with recent theories that question the proposed equivalence of positive/negative affect with approach/avoidance motivation (see, e.g., Eder et al., 2013; Gable & Harmon-Jones, 2010; Harmon-Jones, 2018). Specifically, these theories propose that the approach/avoidance tendencies of different emotions are distinct from (though possibly correlated with) their positive/negative valence, and point to the example of anger as a negatively valenced emotion that is nevertheless associated with approach rather than avoidance behavior (Carver & Harmon-Jones, 2009). Indeed, some preliminary work indicates that experimentally induced anger might be associated with subtle behavioral changes in the reward/punishment Go/No-Go task (Wonderlich, 2020). Conversely, Mkrtchian et al. (2017) showed that experimentally inducing fear/anxiety using a threat-of-shock paradigm increased the strength of aversive Pavlovian influences on instrumental learning of inhibitory actions. Taken together, these results suggest that rather than being linked with the umbrella classes of positive and negative affect—which were the focus of the present study— Pavlovian-instrumental interactions might be more specifically modulated by the approach/avoidance tendencies inherent to specific discrete emotions such as anger and anxiety/fear.

Notwithstanding the null main effect of positive/negative affect, our exploratory analyses suggested that individual-level effects of positive/negative affect on behavior were moderated by several demographic and individual-difference factors, notably including self-reported trait behavioral inhibition (BIS subscale of BIS/BAS). Carver and White (1994) originally proposed the BIS scale as a measure of overall behavioral sensitivity to punishment/aversive stimuli, in line with Gray's (1975) theory of personality. This proposal is partly consistent with the moderating effect that we observed: after the negative affect induction, we did indeed find that participants with higher BIS scores tended to show an increase in the strength of Pavlovian-

instrumental interactions (Pavlovian bias parameter of model). However, inconsistent with theory, we also found that individuals with higher BIS scores showed an *increase* in the strength of their overall behavioral approach tendency (go-bias parameter of model) after a negative affect induction. By contrast, after the positive affect induction, participants with higher BIS scores tended to show a *reduction* in the strength of both behavioral approach tendencies and Pavlovian-instrumental interactions. Overall, these results suggest that higher BIS scores were associated with a strengthening of prepotent response biases by negative affect (i.e., strengthening of both go-biases and Pavlovian biases), whereas positive affect tended to weaken these prepotent biases in those with high BIS scores. This interpretation is akin to documented individual differences in the potentiation of impulsive behavior by negative affect (termed "negative urgency"; Cyders & Smith, 2008; Johnson et al., 2020), although it runs counter to some other behavioral correlates of high BIS scores (cf. Corr et al., 1995; Crockett et al., 2009). Further research is required to tease apart these inconsistencies.

Separately, we also found a moderating effect of gender that was driven by differences between women and men in the effects of the negative affect induction on Pavlovian-instrumental interactions. After a negative affect induction, women tended to show an *increase* in the strength of Pavlovian biases, whereas men showed a *decrease* in Pavlovian bias strength. By contrast, there was no significant effect of the positive affect induction on behavior for either women or men. In interpreting these moderating effects of gender it is important to exercise caution and avoid biological determinism, since participants' self-reported gender is best understood as a categorical proxy for continuous individual differences in many additional biological and social factors (Lindqvist et al., 2020). This caveat notwithstanding, it is noteworthy that one previous study also reported gender differences in the effects of induced

affect on approach/avoidance behavior (Robinson & Sahakian, 2009). Likewise, in a study that solely included women, de Carli et al. (2017) found that behavioral avoidance of sad faces was amplified by a negative affect induction, consistent with the effects of negative affect on women here. It is also noteworthy that women in the present study also showed stronger Pavlovian biases in the first stimulus set, prior to any affect induction (though this contrasts with the findings of Moutoussis et al., 2018, who found no gender differences in behavior in a version of this task). These results suggest that further research is called for to explicate potential gender differences both in Pavlovian-instrumental interactions as well as the modulation of these biases by participants' affective state.

Several limitations of the present study should be noted. First, although our manipulation-check analyses suggested that positive and negative affect inductions were successful in both experiments, we cannot rule out the possibility that this was partially driven by demand effects. This limitation is common to most studies using experimental affect inductions, since it is not difficult for participants to infer the intended effect of the affect induction (Joseph et al., 2020). However, a meta-analysis of previous studies supports the efficacy of affect-induction procedures for online behavioral research (Ferrer et al., 2015), and to ensure data quality in the present study we used a strict set of attention checks—including comprehension questions and detection of browser interactions—to exclude participants who did not attend to the affect-induction videos. Nevertheless, demand effects may still represent a source of unmodelled variance in participants' responses to the affect induction. Moreover, it should be noted that shifts in participants' affect in our study were smaller in magnitude and more transient than shifts of affect that are seen in psychological disorders like major depression. Consequently, our results do not rule out the possibility that Pavlovian-instrumental interactions might shift with larger and

more prolonged shifts in affect. Similarly, the design of our task, with a single set of stimuli per affect induction, meant that the majority of trials were performed while participants were at asymptotic levels of learning. As a result, our task design was not optimized to detect any possible effects of the affect induction on behavior that occurred solely within the initial learning phase over the first 10-15 trials. This question could be addressed in future research by studies using a more specialized task design.

Second, although Experiment 2 was a preregistered replication of Experiment 1, this preregistration applied only to the overall main effects of the affect inductions, and not to the moderation analyses that we discussed above (since we did not have sufficient statistical power in Experiment 1 alone to investigate the moderating effects of individual-difference variables). We attempted to control for the possible inflation of error rates across these moderation analyses by adopting a stricter criterion for identifying credible effects (using 99.5% rather than 95% Bayesian HDIs). However, this constraint does not alter the fact that moderation analyses were exploratory rather than confirmatory; as such, these results should be considered preliminary pending replication in future work.

Overall, our results should be understood in the context of a recent body of computational research into the influences of affect/mood on reinforcement learning (Bennett, Davidson, et al., 2021; Bennett & Niv, 2020; Eldar et al., 2016, 2018; Eldar & Niv, 2015; Neville et al., 2021; Vinckier et al., 2018). One emerging theme from this body of research is the importance of individual differences in the effects of affect on behavior. For instance, similar to the moderating effects of self-reported behavioral inhibition that we observed, Eldar & Niv (2015) previously reported that self-reported trait hypomania moderated the effects of an affect induction on reinforcement learning in a two-armed bandit task. Taken together, these results suggest that it is

crucial for computational theories of affect to consider how individual-difference factors might moderate the effects of affect on instrumental behavior, in addition to any group-level effects.

At the group level, the results of the present study help to specify the effects of affect on reinforcement learning by providing evidence *against* a main effect of positive/negative affect on approach/avoidance behavior or Pavlovian-instrumental interactions. Importantly, however, this does not rule out other hypothesized links, such as the posited influence of affect on value learning for different actions or environmental states (Eldar et al., 2016; Eldar & Niv, 2015), or the proposal that positive (negative) affect directly reinforces (punishes) concurrent actions (Bennett, Davidson, et al., 2021). Further empirical research is required to disentangle these competing theories.

More broadly, the growing set of distinct moderating effects and individual differences that have been documented in the literature call into question whether there is, in fact, a single underlying mechanism that explains all of the effects of affect on reinforcement learning across the population. Instead, in line with the multifaceted effects of affect on other aspects of cognition (see, e.g., Bower, 1981; Clore & Palmer, 2009; Schwarz & Clore, 1983), it may be more plausible that the effects of affective states on reinforcement learning are complex and extensive, interacting with a patchwork of individual differences and contextual factors to shape instrumental behavior.

Acknowledgements

The work presented in this manuscript was supported by the National Institute of Mental Health under award number R01MH119511. DB received salary support from the National Health and Medical Research Council of Australian (fellowship #1165010).

Declaration of interest statement

The authors declare no conflicts of interest.

Reference List

Albrecht, M. A., Waltz, J. A., Cavanagh, J. F., Frank, M. J., & Gold, J. M. (2016). Reduction of Pavlovian bias in schizophrenia: Enhanced effects in clozapine-administered patients. *PLoS One*, *11*(4), e0152781.

Balleine, B. W. (2019). The meaning of behavior: Discriminating reflex and volition in the brain. *Neuron*, *104*(1), 47–62. https://doi.org/10.1016/j.neuron.2019.09.024

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bennett, D., Davidson, G., & Niv, Y. (2021). A model of mood as integrated advantage. *Psychological Review*. https://doi.org/10.1037/rev0000294

Bennett, D., & Niv, Y. (2020). Opening Burton's clock: Psychiatric insights from computational cognitive models. In *The Cognitive Neurosciences (6th. Ed)*. The MIT Press.

Bennett, D., Radulescu, A., Zorowitz, S., Felso, V., & Niv, Y. (2021). *Affect-congruent attention modulates generalized reward expectations* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/vu2cw

Betella, A., & Verschure, P. F. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS One*, *11*(2), e0148037.

Bouhuys, A. L., Bloem, G. M., & Groothuis, T. G. (1995). Induction of depressed and elated mood by music influences the perception of facial emotional expressions in healthy subjects. *Journal of Affective Disorders*, *33*(4), 215–226.

Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*(2), 129–148.

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1990). Startle reflex modification: Emotion or attention? *Psychophysiology*, *27*(5), 513–522. https://doi.org/10.1111/j.1469-8986.1990.tb01966.x

Brown, J. S. (1948). Gradients of approach and avoidance responses and their relation to level of motivation. *Journal of Comparative and Physiological Psychology*, *41*(6), 450–465. https://doi.org/10.1037/h0055463

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, *75*(5), 839–855.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin*, *135*(2), 183–204. https://doi.org/10.1037/a0013965

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333.

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal theta overrides Pavlovian learning biases. *Journal of Neuroscience*, *33*(19), 8541–8548.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. https://doi.org/10.1177/0146167299025002007

Clore, G. L., & Palmer, J. (2009). Affective guidance of intelligent agents: How emotion controls cognition. *Cognitive Systems Research*, *10*(1), 21–30.

Corr, P. J., Pickering, A. D., & Gray, J. A. (1995). Personality and reinforcement in associative and instrumental learning. *Personality and Individual Differences*, *19*(1), 47–71.

Crockett, M. J., Clark, L., & Robbins, T. W. (2009). Reconciling the role of serotonin in behavioral inhibition and aversion: Acute tryptophan depletion abolishes punishment-induced inhibition in humans. *Journal of Neuroscience*, *29*(38), 11993–11999.

Csifcsák, G., Melsæter, E., & Mittner, M. (2020). Intermittent absence of control during reinforcement learning interferes with Pavlovian bias in action selection. *Journal of Cognitive Neuroscience*, *32*(4), 646–663.

Cyders, M. A., & Smith, G. T. (2008). Emotion-based dispositions to rash action: Positive and negative urgency. *Psychological Bulletin*, *134*(6), 807–828.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*(2), 285–298.

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8), 1153–1160.

De Carli, P., Riem, M. M. E., & Parolin, L. (2017). Approach-avoidance responses to infant facial expressions in nulliparous women: Associations with early experience and mood induction. *Infant Behavior and Development*, *49*, 104–113. https://doi.org/10.1016/j.infbeh.2017.08.005

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Dickinson, A., & Dearing, M. F. (1979). Appetitive-aversive interactions and inhibitory processes. In *Mechanisms of Learning and Motivation: A Memorial Volume to Jerzy Konorski* (pp. 203–231).

Dorfman, H. M., & Gershman, S. J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. *Nature Communications*, *10*(1), 5826. https://doi.org/10.1038/s41467-019-13737-7

Eder, A. B., Elliot, A. J., & Harmon-Jones, E. (2013). Approach and Avoidance Motivation: Issues and Advances. *Emotion Review*, *5*(3), 227–229. https://doi.org/10.1177/1754073913477990

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*, 6149.

Eldar, E., Roth, C., Dayan, P., & Dolan, R. J. (2018). Decodability of reward learning signals predicts mood fluctuations. *Current Biology*, *28*(9), 1433–1439.

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24.

Eliot, A. J. (2008). *Handbook of Approach and Avoidance Motivation*. Psychology Press.

Elliot, A. J., Eder, A. B., & Harmon-Jones, E. (2013). Approach–Avoidance Motivation and Emotion: Convergence and Divergence. *Emotion Review*, *5*(3), 308–311. https://doi.org/10.1177/1754073913477517

Ferrer, R. A., Grenen, E. G., & Taber, J. M. (2015). Effectiveness of internet-based affect induction procedures: A systematic review and meta-analysis. *Emotion*, *15*(6), 752–762.

Gable, P., & Harmon-Jones, E. (2010). The motivational dimensional model of affect: Implications for breadth of attention, memory, and cognitive categorisation. *Cognition and Emotion*, *24*(2), 322–337.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Geurts, D. E., Huys, Q. J., Den Ouden, H. E., & Cools, R. (2013). Aversive Pavlovian control of instrumental behavior in humans. *Journal of Cognitive Neuroscience*, *25*(9), 1428–1441.

Gray, J. A. (1975). *Elements of a Two-Process Theory of Learning*. Academic Press.

Green, P., & MacLeod, C. J. (2016). simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J., Dayan, P., Dolan, R. J., & Duzel, E. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *Journal of Neuroscience*, *31*(21), 7867–7875.

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage*, *62*(1), 154–166.

Harmon-Jones, E. (2018). On motivational influences, moving beyond valence, and integrating dimensional and discrete views of emotion. *Cognition and Emotion*, 1–8. https://doi.org/10.1080/02699931.2018.1514293

Hershberger, W. A. (1986). An approach through the looking-glass. *Animal Learning & Behavior*, *14*(4), 443–451.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280–1300.

Johnson, S. L., Elliott, M. V., & Carver, C. S. (2020). Impulsive responses to positive and negative emotions: Parallel neurocognitive correlates and their implications. *Biological Psychiatry*, *87*(4), 338–349.

Joseph, D. L., Chan, M. Y., Heintzelman, S. J., Tay, L., Diener, E., & Scotney, V. S. (2020). The manipulation of affect: A meta-analysis of affect induction procedures. *Psychological Bulletin*. https://doi.org/10.1037/bul0000224

Kasch, K. L., Rottenberg, J., Arnow, B. A., & Gotlib, I. H. (2002). Behavioral activation and inhibition systems and the severity and course of depression. *Journal of Abnormal Psychology*, *111*(4), 589–597.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kertz, S. J., Lee, J., & Björgvinsson, T. (2014). Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychological Assessment*, *26*(4), 1146–1154. https://doi.org/10.1037/a0037251

Konorski, J. (1967). *Integrative Activity of the Brain: An Interdisciplinary Approach*. University of Chicago Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lang, P. J. (1995). Studies of motivation and attention. *American Psychologist*, *50*(5), 372–385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*(3), 377–195.

Lindqvist, A., Sendén, M. G., & Renström, E. A. (2020). What is gender, anyway: A review of the options for operationalising gender. *Psychology & Sexuality*, 1–13. https://doi.org/10.1080/19419899.2020.1729844

McNaughton, N. (1989). *Biology and Emotion*. Cambridge University Press.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Meyer, B., Beevers, C. G., Johnson, S. L., & Simmons, E. (2007). Unique association of approach motivation and mania vulnerability. *Cognition & Emotion*, *21*(8), 1647–1668. https://doi.org/10.1080/02699930701252686

Meyer, B., Johnson, S. L., & Winters, R. (2001). Responsiveness to threat and incentive in bipolar disorder: Relations of the BIS/BAS scales with symptoms. *Journal of Psychopathology and Behavioral Assessment*, *23*(3), 133–143.

Millner, A. J., Gershman, S. J., Nock, M. K., & den Ouden, H. E. (2018). Pavlovian control of escape and avoidance. *Journal of Cognitive Neuroscience*, *30*(10), 1379–1390.

Mkrtchian, A., Roiser, J. P., & Robinson, O. J. (2017). Threat of shock and aversive inhibition: Induced anxiety modulates Pavlovian-instrumental interactions. *Journal of Experimental Psychology: General*, *146*(12), 1694–1704. https://doi.org/10.1037/xge0000363

Moutoussis, M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., Dolan, R. J., Dayan, P., & on behalf of The Neuroscience in Psychiatry Network Research Consortium. (2018). Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. *PLOS Computational Biology*, *14*(12), e1006679. https://doi.org/10.1371/journal.pcbi.1006679

Neville, V., Dayan, P., Gilchrist, I. D., Paul, E. S., & Mendl, M. (2021). Dissecting the links between reward and loss, decision-making, and self-reported affect using a computational approach. *PLOS Computational Biology*, *17*(1), e1008555. https://doi.org/10.1371/journal.pcbi.1008555

Ophir, Y., Sisso, I., Asterhan, C. S., Tikochinski, R., & Reichart, R. (2020). The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. *Clinical Psychological Science*, *8*(1), 65–83.

Pagliaccio, D., Luking, K. R., Anokhin, A. P., Gotlib, I. H., Hayden, E. P., Olino, T. M., Peng, C.-Z., Hajcak, G., & Barch, D. M. (2016). Revising the BIS/BAS Scale to study development: Measurement invariance and normative effects of age and sex from childhood through adulthood. *Psychological Assessment*, *28*(4), 429–442. https://doi.org/10.1037/pas0000186

Raab, H. A., & Hartley, C. A. (2020). Adolescents exhibit reduced Pavlovian biases on instrumental learning. *Scientific Reports*, *10*(1), 15770. https://doi.org/10.1038/s41598-020-72628-w

Robinson, O. J., & Sahakian, B. J. (2009). A double dissociation in the roles of serotonin and mood in healthy subjects. *Biological Psychiatry*, *65*(1), 89–92. https://doi.org/10.1016/j.biopsych.2008.10.001

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513.

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A scale for the assessment of hedonic tone: The Snaith-Hamilton Pleasure Scale. *British Journal of Psychiatry*, *167*(1), 99–103. https://doi.org/10.1192/bjp.167.1.99

Solarz, K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology*, *59*(4), 239–245.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Swart, J. C., Frank, M. J., Määttä, J. I., Jensen, O., Cools, R., & den Ouden, H. E. (2018). Frontal network dynamics reflect neurocomputational mechanisms for reducing maladaptive biases in motivated action. *PLoS Biology*, *16*(10), e2005979.

Swart, J. C., Froböse, M. I., Cook, J. L., Geurts, D. E., Frank, M. J., Cools, R., & Den Ouden, H. E. (2017). Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *ELife*, *6*, e22169.

Trew, J. L. (2011). Exploring the roles of approach and avoidance in depression: An integrative model. *Clinical Psychology Review*, *31*(7), 1156–1168.

Urošević, S., Abramson, L. Y., Harmon-Jones, E., & Alloy, L. B. (2008). Dysregulation of the behavioral approach system (BAS) in bipolar spectrum disorders: Review of theory and evidence. *Clinical Psychology Review*, *28*(7), 1188–1205.

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2021). Bayes factors for mixed models. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-021-00113-2

Vinckier, F., Rigoux, L., Oudiette, D., & Pessiglione, M. (2018). Neuro-computational account of how mood fluctuations arise and affect decision making. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03774-z

Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of Abnormal Psychology*, *97*(4), 487–491.

Vrijsen, J. N., van Oostrom, I., Speckens, A., Becker, E. S., & Rinck, M. (2013). Approach and avoidance of emotional faces in happy and sad mood. *Cognitive Therapy and Research*, *37*(1), 1–6. https://doi.org/10.1007/s10608-012-9436-9

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, *75*(5), 820–838.

Wonderlich, J. A. (2020). *Anger and Pavlovian Bias: Integrating Laboratory Task Performance and Ecological Momentary Assessment* [Doctor of Philosophy thesis, George Mason University]. https://www.proquest.com/openview/e8273333f43823d72ddd76ae239d3241

Youngstrom, E. A., Murray, G., Johnson, S. L., & Findling, R. L. (2013). The 7 Up 7 Down Inventory: A 14-item measure of manic and depressive tendencies carved from the General Behavior Inventory. *Psychological Assessment*, *25*(4), 1377–1383. https://doi.org/10.1037/a0033975

Zorowitz, S., Niv, Y., & Bennett, D. (2021). *Inattentive responding can induce spurious associations between task behavior and symptom measures* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/rynhk

Supplementary Material for

The effects of induced positive and negative affect on Pavlovian-instrumental interactions in learning

Isla Weber[1], Sam Zorowitz[1], Yael Niv[1,2], Daniel Bennett[3]

1 Princeton Neuroscience Institute, Princeton University, USA

2 Department of Psychology, Princeton University, USA

3 Department of Psychiatry, Monash University, Australia

Section S1: Task Instructions

The following text was formatted using HTML and CSS for presentation to participants within their web browser:

*In this task, you will be inspecting robots as they move down the assembly line into the **scanner**. When a robot is scanned, you must decide whether to*

- ***Approve** a robot as complete (press SPACE)*
- ***Reject** a robot as incomplete (do nothing)*

*Next we will practice these actions. Four robots will come down the assembly line. **Approve** each robot by pressing SPACE. HINT: only press once the robot is in the scanner and the scanner light comes on.*

< 4 practice trials >

*Great job! Four more robots will come down the assembly line. **Reject** each robot by doing nothing.*

< 4 practice trials >

During the scanner, the scanner will shine one of two colours. If the scanner is blue/yellow [randomized across participants], you will earn +10 points for correctly judging the robot as complete or incomplete. Incorrect judgments will earn you +1 points.

If the scanner is yellow/blue [randomized across participants], you will earn -1 points for correctly judging the robot as complete or incomplete. Incorrect judgments will earn you -10 points.

*When a robot is scanned, it will reveal a **symbol** on its chestplate. This symbol will mark whether a robot is complete or incomplete. Pay close attention to the symbol as it will help you decide whether to accept (press SPACE) or reject (do nothing) the robot.*

Be aware: sometimes the scanner will malfunction and provide you incorrect feedback. That is, it may provide you the wrong number of points for your judgment.

At the end of the task, the total number of points you've earned will be converted into a performance bonus.

Next, we will ask you some questions about the task.

The questions below were used to check participants' comprehension of instructions prior to their proceeding to the task. Participants were unable to proceed until they had answered all questions correctly.

Q1: To reject a robot (i.e., judge as incomplete), what do you do?

(A) Press SPACE *(B) Do nothing* *(C) Press enter*

Correct answer: **(B)**

Q2: When the scanner light is [randomized gain-domain colour], how many points will you earn for a correct judgment?

(A) +10 *(B) +1* *(C) -1* *(D) -10*

Correct answer: **(A)**

Q3: When the scanner light is [randomized loss-domain colour], how many points will you earn for a correct judgment?

(A) +10 *(B) +1* *(C) -1* *(D) -10*

Correct answer: **(C)**

Q4: True or False: the scanner will sometimes malfunction and provide incorrect feedback.

(A) True *(B) False*

Correct answer: **(A)**

Q5: Will the number of points you earn affect your performance bonus?

(A) Yes *(B) No*

Correct answer: **(A)**

Section S2: Mixed-effects regression tables, Experiment 1

Analysis 1: Effects of affect induction on change in mood valence

Dependent variable: change in self-reported mood valence (post-video valence minus pre-video valence)

Table S1a. Overview of regression analysis 1

Fixed effects	**Participant-wise random effects**
- Intercept	- Random intercept
- Video condition	- Random slope for block number
- Block number	
- Video condition * block number	

Table S1b. Coefficient estimates and inferential statistics for regression analysis 1

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	19.05 (1)	0.01	< .001	***
Video condition	-	120.60 (2)	-	< .001	***
-	Positive (vs. neutral)	-	0.10	.02	*
-	Negative (vs. neutral)	-	-0.26	< .001	***
Block	-	0.38 (1)	0.01	.54	
Video condition * Block	-	1.73 (2)	-	.42	
-	Positive * Block (vs. neutral)	-	-0.02	.40	
-	Negative * Block (vs. neutral)	-	0.01	.86	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 2: Effects of affect induction on change in mood arousal, Experiment 1

Dependent variable: change in self-reported mood arousal (post-video arousal minus pre-video arousal)

Table S2a. Overview of regression analysis 2

Fixed effects	**Participant-wise random effects**
- Intercept	- Random intercept
- Video condition	- Random slope for block number
- Block number	
- Video condition * block number	

Table S2b. Coefficient estimates and inferential statistics for regression analysis 2

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	2.41 (1)	-0.06	.12	
Video condition	-	6.35 (2)	-	.04	*
-	Positive (vs. neutral)	-	0.09	.03	*
-	Negative (vs. neutral)	-	0.03	.50	
Block	-	0.02 (1)	-0.003	.87	
Video condition * Block	-	0.69 (2)	-	.71	
-	Positive * Block (vs. neutral)	-	0.01	.63	
-	Negative * Block (vs. neutral)	-	0.02	.41	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 3: Analysis of choice behavior (without affect-related regressors), Experiment 1

Dependent variable: Go/No-Go decision accuracy (coded as incorrect: 0, correct: 1).

Table S3a. Overview of regression analysis 3

Fixed effects	**Participant-wise random effects**
- Intercept	- Random intercept
- Block number	- Random slopes for:
- Set number	- Block number
- Outcome domain	- Set number
- Preferred action	- Outcome domain
- Outcome domain * preferred action	- Preferred action
- Set number * Block number	- Outcome domain * Preferred action
- Set number * Outcome domain	- Set number * Block number
- Set number * Preferred action	- Set number * Outcome domain
- Set number * Outcome domain * Preferred action	- Set number * Preferred action
	- Set number * Outcome domain * Preferred action

Table S3b. Coefficient estimates and inferential statistics for regression analysis 3

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	35.02 (1)	0.84	< .001	***
Block number	-	114.59 (1)	1.15	< .001	***
Set number	-	17.97 (1)	0.82	< .001	***
Outcome domain	Gain (vs. loss)	65.86 (1)	2.03	< .001	***
Preferred action	No-Go (vs. Go)	0.33 (1)	0.09	.56	
Outcome domain * Preferred action	Gain (vs.loss) * No-Go (vs. Go)	80.55 (1)	-3.49	< .001	***
Set number * Block number	-	1.92 (1)	0.19	.17	
Set number * Outcome domain	Set number * Gain (vs.loss)	4.07 (1)	-0.62	.04	*
Set number * Preferred action	Set number * No-Go (vs. Go)	1.45 (1)	-0.30	.23	
Set number * Outcome domain * Preferred action	Set number * Gain (vs.loss) * No-Go (vs. Go)	4.79 (1)	0.90	.03	*

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 4: Analysis of choice behavior (with affect-related regressors), Experiment 1

Dependent variable: Go/No-Go decision accuracy (coded as incorrect: 0, correct: 1).

Table S4a. Overview of regression analysis 4

Fixed effects	**Participant-wise random effects**

- Intercept
- Block number
- Set number
- Outcome domain
- Preferred action
- Affect condition
- Outcome domain * preferred action
- Set number * Block number
- Set number * Outcome domain
- Set number * Preferred action
- Set number * Affect condition
- Set number * Outcome domain * Preferred action
- Set number * Outcome domain * Affect condition
- Set number * Preferred action * Affect condition
- Set number * Outcome domain * Preferred action * Affect condition
	- Random intercept
	- Random slopes for:
	  - Block number
	  - Set number
	  - Outcome domain
	  - Preferred action
	  - Outcome domain * Preferred action
	  - Set number * Block number
	  - Set number * Outcome domain
	  - Set number * Preferred action
	  - Set number * Outcome domain * Preferred action

Table S4b. Coefficient estimates and inferential statistics for regression analysis 4

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	32.33 (1)	0.93	< .001	***
Block number	-	114.28 (1)	1.15	< .001	***
Set number	-	6.97 (1)	0.67	< .01	**
Outcome domain	Gain (vs. loss)	65.84 (1)	2.02	< .001	***
Preferred action	No-Go (vs. Go)	0.34 (1)	0.09	.56	
Affect condition	Negative affect (vs. positive)	1.12 (1)	-0.16	.29	
Outcome domain * Preferred action	Gain (vs.loss) * No-Go (vs. Go)	80.71(1)	-3.48	< .001	***
Set number * Block number	-	1.83 (1)	0.19	.18	
Set number * Outcome domain	Set number * Gain (vs.loss)	0.76 (1)	-0.32	.38	
Set number * Preferred action	Set number * No-Go (vs. Go)	0.46 (1)	-0.23	.50	
Set number * Affect condition	Set number * Negative affect (vs. positive)	0.89 (1)	0.30	.35	
Set number * Outcome domain * Preferred action	Set number * Gain (vs.loss) * No-Go (vs. Go)	1.22 (1)	0.60	.27	
Set number * Outcome domain * Affect condition	Set number * Gain (vs.loss) * Negative affect (vs. positive)	2.38 (1)	-0.59	.12	
Set number * Preferred action * Affect condition	Set number * No-Go (vs. Go) * Negative affect (vs. positive)	0.08 (1)	-0.13	.77	
Set number * Outcome domain * Preferred action * Affect condition	Set number * Gain (vs.loss) * No-Go (vs. Go) * Negative affect (vs. positive)	0.81 (1)	0.60	.37	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 5: Correlations between self-report measures and affect-related random slopes

Table S5a. Correlation matrix of self-report measures and positive-affect-related random slopes (N = 44)

Random slope	Hypomania (7-up)	Depression (7-down)	BIS	BAS (reward sensitivity)	BAS (drive)	Age
Set number	.05 (.86)	-.29 (.18)	-.34 (.09)	.16 (.52)	.13 (.66)	-.20 (.38)
Set number * Outcome domain	-.04 (.88)	.22 (.37)	.30 (.16)	-.11 (.71)	.01 (.97)	.07 (.82)
Set number * Preferred action	.03 (.91)	.06 (.85)	.16 (.52)	.09 (.80)	.13 (.69)	-.03 (.91)
Set number * Outcome domain * Preferred action	.07 (.82)	-.21 (.37)	-.29 (.18)	.07 (.82)	.12 (.71)	-.05 (.86)

Spearman rank-order correlations, false-discovery-rate corrected at α = .05. Corrected p-values for each correlation are reported in parentheses.

Table S5b. Correlation matrix of self-report measures and negative-affect-related random slopes (N = 47)

Random slope	Hypomania (7-up)	Depression (7-down)	BIS	BAS (reward sensitivity)	BAS (drive)	Age
Set number	.04 (.98)	-.07 (.98)	-.04 (.98)	< .01 (.99)	< .01 (.99)	.01 (.99)
Set number * Outcome domain	-.10 (.87)	.10 (.87)	-.18 (.61)	-.10 (.87)	-.03 (.98)	-.12 (.85)
Set number * Preferred action	-.12 (.86)	< .01 (.99)	-.18 (.61)	-.08 (.89)	-.09 (.89)	-.12 (.85)
Set number * Outcome domain * Preferred action	.20 (.56)	-.16 (.69)	.06 (.98)	.05 (.98)	.10 (.87)	.02 (.98)

Spearman rank-order correlations, false-discovery-rate corrected at α = .05. Corrected p-values for each correlation are reported in parentheses.

Section S3: Effects of affect induction on self-reported arousal

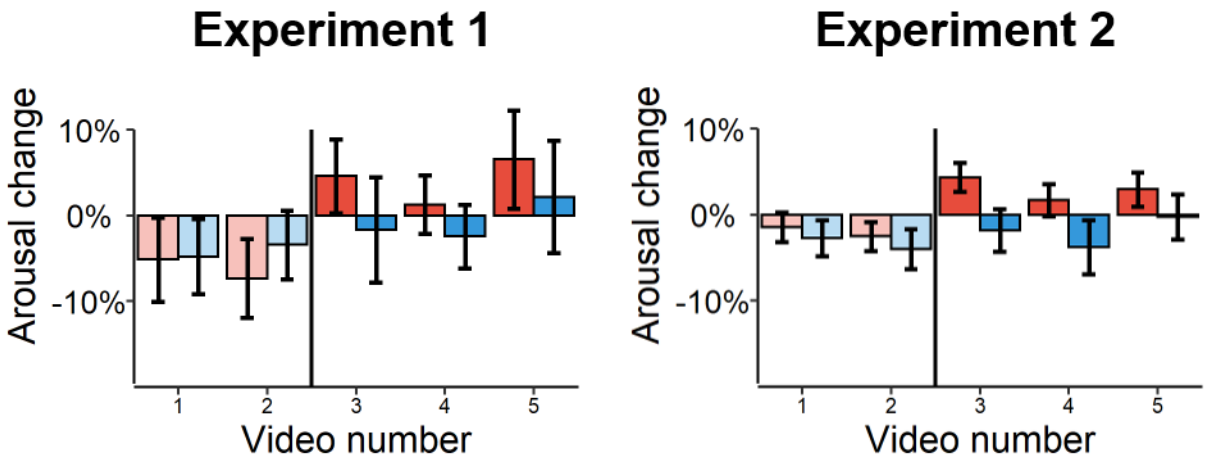


Figure S1. Mean change in mood arousal produced by each video (post-video minus pre-video arousal as a percentage of the total length of the Affective Slider) for Experiment 1 (left panel) and Experiment 2 (right panel). Within each panel, the solid vertical line demarcates the boundary between stimulus set one (left, in which only neutral videos were presented) and set two (right, in which participants viewed either three happy or three sad videos. Dark red and dark blue bars denote happy and sad videos, respectively. Pale bars for videos 1 and 2 denote responses to neutral videos by participants who subsequently viewed happy videos (pale red) or subsequently viewed sad videos (pale blue). There was significant effect of the affect induction on self-reported mood arousal, $\chi^2(2) = 6.34$, $p < .05$, driven by a tendency for happy videos to increase arousal more than neutral videos, ($\beta = 0.09$, $p < .05$). There was no difference between arousal change following neutral videos and arousal change following sad videos ($\beta = 0.03$, $p = .50$). Error bars represent the 95% confidence interval of the mean.

Section S4: Details of participant recruitment for Experiment 2

General procedure. Participants were recruited through Prolific using a two-stage process. First, we collected questionnaire and demographic data from adult participants residing in the United States and Canada, obtaining a sample larger than our intended final target sample size (N = 1190). In this stage, psychological self-report measures were presented in a counterbalanced order across participants. All survey participants who passed questionnaire-based screening (N = 890; see below for details) were then given the opportunity to complete the task portion of the study within 7 days of completing the questionnaires. The task was identical to that reported in Experiment 1. Based on anticipated rates of participant exclusion (25%, as per Experiment 1), for a target sample of > 300, we collected task data from 400 participants. This study was approved by the Institutional Review Board of Princeton University, and all participants provided written informed consent. Total study duration was approximately 35 minutes per participant. To ensure that compensation for questionnaire completion was not dependent on task completion, participants received partial monetary compensation at the conclusion of the questionnaire stage (USD$1 for the questionnaire screening, which took approximately 5 minutes). As in Experiment 1, participants who completed the behavioral task received a task payment that included an incentive-compatible bonus for task performance (mean total task payment = USD $5.49, SD = 0.31).

Participants. Initial questionnaire data were collected from 1190 participants. As a data quality-assurance measure, questionnaires each included one *infrequency item* that could be used to screen for inattentive responding (e.g., selecting responses at random). Participants who gave implausible or impossible responses on any of these screening measures (e.g., endorsing a survey item stating that they are able to go several days at a time without breathing) were not invited back to participate in the task phase of the experiment. 300 participants (25.2%) gave a problematic response to one or more infrequency items, and were not invited back to complete the task portion of the study. Among the participants who passed questionnaire screening, 400 subsequently completed the behavioral task. Of these, 65 participants (16.25% of behavioral sample) were excluded after application of our preregistered exclusion criteria (identical to those for Experiment 1). The resulting sample size exceeded our target sample size of 300, and so all 335 participants were included in analyses.

Section S5: Mixed-effects regression tables, Experiment 2

Analysis 1: Effects of affect induction on change in mood valence, Experiment 2

Dependent variable: change in self-reported mood valence (post-video valence minus pre-video valence)

Table S6a. Overview of regression analysis 1

Fixed effects	**Participant-wise random effects**
- Intercept	- Random intercept
- Video condition	- Random slope for block number
- Block number	
- Video condition * block number	

Table S6b. Coefficient estimates and inferential statistics for regression analysis 1

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	0.76 (1)	-0.01	.39	
Video condition	-	308.25 (2)	-	< .001	***
-	Positive (vs. neutral)	-	0.13	< .001	***
-	Negative (vs. neutral)	-	-0.18	< .001	***
Block	-	2.03 (1)	-0.02	.15	
Video condition * Block	-	6.14 (2)	-	.046	*
-	Positive * Block (vs. neutral)	-	0.02	.20	
-	Negative * Block (vs. neutral)	-	0.04	.01	*

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 2: Effects of affect induction on change in mood arousal, Experiment 2

Dependent variable: change in self-reported mood arousal (post-video arousal minus pre-video arousal)

Table S7a. Overview of regression analysis 2

Fixed effects	**Participant-wise random effects**
- Intercept	- Random intercept
- Video condition	- Random slope for block number
- Block number	
- Video condition * block number	

Table S7b. Coefficient estimates and inferential statistics for regression analysis 2

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	6.67 (1)	-0.04	.01	**
Video condition	-	23.71 (2)	-	< .001	***
-	Positive (vs. neutral)	-	0.08	< .001	***
-	Negative (vs. neutral)	-	0.02	.33	
Block	-	1.20 (1)	-0.003	.27	
Video condition * Block	-	2.92 (2)	-	.23	
-	Positive * Block (vs. neutral)	-	0.005	.72	
-	Negative * Block (vs. neutral)	-	0.02	.14	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 3: Analysis of choice behavior (without affect-related regressors), Experiment 2

Dependent variable: Go/No-Go decision accuracy (coded as incorrect: 0, correct: 1).

Table S8a. Overview of regression analysis 3

Fixed effects	**Participant-wise random effects**
- Intercept	- Random intercept
- Block number	- Random slopes for:
- Set number	- Block number
- Outcome domain	- Set number
- Preferred action	- Outcome domain
- Outcome domain * preferred action	- Preferred action
- Set number * Block number	- Outcome domain * Preferred action
- Set number * Outcome domain	- Set number * Block number
- Set number * Preferred action	- Set number * Outcome domain
- Set number * Outcome domain * Preferred action	- Set number * Preferred action
	- Set number * Outcome domain * Preferred action

Table S8b. Coefficient estimates and inferential statistics for regression analysis 3

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	78.96 (1)	0.65	< .001	***
Block number	-	341 (1)	0.95	< .001	***
Set number	-	25.83 (1)	0.47	< .001	***
Outcome domain	Gain (vs. loss)	244.85 (1)	1.80	< .001	***
Preferred action	No-Go (vs. Go)	2.69 (1)	0.17	.10	
Outcome domain * Preferred action	Gain (vs.loss) * No-Go (vs. Go)	222.37 (1)	-3.23	< .001	***
Set number * Block number	-	26.52 (1)	0.31	< .001	***
Set number * Outcome domain	Set number * Gain (vs.loss)	3.01 (1)	-0.24	.08	
Set number * Preferred action	Set number * No-Go (vs. Go)	2.89 (1)	0.18	.09	
Set number * Outcome domain * Preferred action	Set number * Gain (vs.loss) * No-Go (vs. Go)	0.73 (1)	0.16	.39	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 4: Analysis of choice behavior (with affect-related regressors), Experiment 2

Dependent variable: Go/No-Go decision accuracy (coded as incorrect: 0, correct: 1).

Table S9a. Overview of regression analysis 4

Fixed effects	**Participant-wise random effects**

- Intercept
- Block number
- Set number
- Outcome domain
- Preferred action
- Affect condition
- Outcome domain * preferred action
- Set number * Block number
- Set number * Outcome domain
- Set number * Preferred action
- Set number * Affect condition
- Set number * Outcome domain * Preferred action
- Set number * Outcome domain * Affect condition
- Set number * Preferred action * Affect condition
- Set number * Outcome domain * Preferred action * Affect condition
	- Random intercept
	- Random slopes for:
	  - Block number
	  - Set number
	  - Outcome domain
	  - Preferred action
	  - Outcome domain * Preferred action
	  - Set number * Block number
	  - Set number * Outcome domain
	  - Set number * Preferred action
	  - Set number * Outcome domain * Preferred action

Table S9b. Coefficient estimates and inferential statistics for regression analysis 4

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	61.39 (1)	0.65	< .001	***
Block number	-	340.84 (1)	0.95	< .001	***
Set number	-	16.35 (1)	0.51	< .001	***
Outcome domain	Gain (vs. loss)	244.51 (1)	1.80	< .001	***
Preferred action	No-Go (vs. Go)	2.68 (1)	0.17	.10	
Affect condition	Negative affect (vs. positive)	0.01 (1)	-0.01	.90	
Outcome domain * Preferred action	Gain (vs.loss) * No-Go (vs. Go)	222.11 (1)	-3.23	< .001	***
Set number * Block number	-	26.44 (1)	0.31	< .001	***
Set number * Outcome domain	Set number * Gain (vs.loss)	1.41 (1)	-0.21	.24	
Set number * Preferred action	Set number * No-Go (vs. Go)	0.58 (1)	-0.12	.45	
Set number * Affect condition	Set number * Negative affect (vs. positive)	0.20 (1)	-0.07	.66	
Set number * Outcome domain * Preferred action	Set number * Gain (vs.loss) * No-Go (vs. Go)	0.83 (1)	0.24	.36	
Set number * Outcome domain * Affect condition	Set number * Gain (vs.loss) * Negative affect (vs. positive)	0.09 (1)	-0.06	.77	
Set number * Preferred action * Affect condition	Set number * No-Go (vs. Go) * Negative affect (vs. positive)	0.41 (1)	0.12	.52	
Set number * Outcome domain * Preferred action * Affect condition	Set number * Gain (vs.loss) * No-Go (vs. Go) * Negative affect (vs. positive)	0.16 (1)	-0.13	.69	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Analysis 5: Correlations between self-report measures and affect-related random slopes

Table S10a. Correlation matrix of self-report measures and positive-affect-related random slopes (N = 158)

Random slope	Hypomania (7-up)	Depression (7-down)	BIS	BAS (reward sensitivity)	BAS (drive)	Age
Set number	-.12 (.26)	-.07 (.50)	.13 (.22)	< .01 (.99)	-.13 (.22)	-.14 (.15)
Set number * Outcome domain	.12 (.22)	-.03 (.77)	***-.23 (.02)***	-.03 (.77)	.04 (.74)	***.19 (.04)***
Set number * Preferred action	.08 (.48)	.11 (.32)	-.08 (.48)	< .01 (.99)	.08 (.47)	< .01 (.99)
Set number * Outcome domain * Preferred action	-.08 (.48)	.06 (.52)	.16 (.10)	.09 (.45)	.04 (.70)	-.18 (.06)

Spearman rank-order correlations, false-discovery-rate corrected at $\alpha = .05$. Corrected p-values for each correlation are reported in parentheses. Bold italic typeface denotes significant correlations at $p < .05$ (corrected).

Table S10b. Correlation matrix of self-report measures and negative-affect-related random slopes (N = 177)

Random slope	Hypomania (7-up)	Depression (7-down)	BIS	BAS (reward sensitivity)	BAS (drive)	Age
Set number	-.07 (.71)	.05 (.79)	.05 (.79)	.02 (.89)	-.07 (.69)	-.04 (.79)
Set number * Outcome domain	< .01 (.95)	-.09 (.60)	.01 (.89)	-.01 (.89)	.08 (.63)	.03 (.83)
Set number * Preferred action	.11 (.39)	-.05 (.79)	-.07 (.69)	-.02 (.89)	.05 (.79)	-.04 (.79)
Set number * Outcome domain * Preferred action	-.03 (.84)	.05 (.79)	-.05 (.79)	-.01 (.89)	-.04 (.79)	.04 (.80)

Spearman rank-order correlations, false-discovery-rate corrected at $\alpha = .05$. Corrected p-values for each correlation are reported in parentheses.

Analysis 6: Analysis of choice behavior including effect of light color, Experiment 2

Dependent variable: Go/No-Go decision accuracy (coded as incorrect: 0, correct: 1).

Table S11a. Overview of regression analysis 6

Fixed effects	**Participant-wise random effects**

- Intercept
- Gain-domain color
- Block number
- Set number
- Outcome domain
- Preferred action
- Outcome domain * Preferred action
- Gain-domain color * Outcome domain
- Gain-domain color * Preferred action
- Set number * Block number
- Set number * Outcome domain
- Set number * Preferred action
- Gain-domain color * Outcome domain * Preferred action
- Set number * Outcome domain * Preferred action
	- Random intercept
	- Random slopes for:
	  - Block number
	  - Set number
	  - Outcome domain
	  - Preferred action
	  - Outcome domain * Preferred action
	  - Set number * Block number
	  - Set number * Outcome domain
	  - Set number * Preferred action
	  - Set number * Outcome domain * Preferred action

Table S11b. Coefficient estimates and inferential statistics for regression analysis 6

Omnibus effect	Contrast	χ^2 (df)	ß	*p*	
Intercept	-	54.25 (1)	0.72	< .001	***
Gain-domain color	Yellow (vs. blue)	0.99 (1)	-0.13	.32	
Block number	-	341.26 (1)	0.95	< .001	***
Set number	-	25.46 (1)	0.47	< .001	***
Outcome domain	Gain (vs. loss)	139.91 (1)	1.74	< .001	***
Preferred action	No-Go (vs. Go)	2.39 (1)	0.22	.12	
Outcome domain * Preferred action	Gain (vs.loss) * No-Go (vs. Go)	137.57 (1)	-3.28	< .001	***
Gain-domain color * Outcome domain	Yellow (vs. blue) * Gain (vs. loss)	0.30 (1)	0.10	.58	
Gain-domain color * Preferred action	Yellow (vs. blue) * No-Go (vs. Go)	0.25 (1)	-0.10	.61	
Set number * Block number	-	26.40 (1)	0.31	< .001	***
Set number * Outcome domain	Set number * Gain (vs.loss)	2.94 (1)	-0.24	.09	
Set number * Preferred action	Set number * No-Go (vs. Go)	2.97 (1)	0.19	.08	
Gain-domain color * Outcome domain * Preferred action	Yellow (vs. blue) * Gain (vs.loss) * No-Go (vs. Go)	0.07 (1)	0.10	.79	
Set number * Outcome domain * Preferred action	Set number * Gain (vs.loss) * No-Go (vs. Go)	0.70 (1)	0.16	.40	

Note: * $p < .05$; ** $p < .01$; *** $p < .001$. For rows where both omnibus tests and coefficient estimates are reported, reported *p*-values are for omnibus tests.

Section S6: Overview of computational modelling methods

Overview of models

For computational modelling analyses, we combined participant samples from both experiments (total *N* included in modelling analyses = 426). We fit three models to participants' behavior. All models shared the same underlying structure and solely differed in terms of the structure of the group-level parameter distributions from which each participant's individual-level parameters were assumed to be drawn (detailed below). The common structure of all models was based on a standard model of this task formulated by Guitart-Masip et al. (2012; see also Dayan et al., 2006). This model assumes that, on each trial, participants probabilistically select either a 'Go' action or a 'No-Go' action according to a softmax (logistic) function of each action's *action weight* (denoted $W(s_t, a)$, where $a \in \{Go, NoGo\}$):

$$Pr(choice = Go) = \frac{1}{1 + e^{\beta \times (W(s_t,NoGo) - W(s_t,Go))}} \quad (1)$$

In Equation 1, β is a softmax inverse-temperature parameter that captures the stochasticity of participants' choices (random choices as $\beta \to 0$, deterministic choices of the action with the higher weight as $\beta \to \infty$).

The state s_t represents the rune symbol (and corresponding light color) presented on trial *t*. As such, there were eight discrete states in total across this task: four in the first stimulus set (corresponding to the four different robots/trial types), and another four in the second stimulus set. Each of these states itself has a valence, denoted $V(s_t)$ that depends on whether its payouts were in the gain domain or the loss domain:

$$V(s_t) = \begin{cases} 1, & \text{stimulus} = \text{GW or NGW} \\ -1, & \text{stimulus} = \text{GAL or NGAL} \end{cases} \quad (2)$$

Since in our task the payout domain of each stimulus was instructed to participants (via the scanner light color) rather than learned, we treated the valence of states as static, and did not model a state-valence learning process (Millner et al., 2018; Swart et al., 2017).

As shown in Equation 3, the action weights of each action in each state were assumed to be composed of three components: (i) a learned action-value $Q(s_t, a)$, (ii) a generalized go-bias (controlled by a parameter *b*) that reflects the overall tendency of participants to prefer making Go responses $(b > 0)$ or No-Go responses $(b < 0)$, and (iii) a Pavlovian-instrumental interaction component (controlled by the parameter π) that reflects the tendency (where $\pi > 0$) for participants to prefer Go responses in appetitive states and No-Go responses in aversive states (note that since the softmax subtracts the weights of the two possible actions, it is immaterial whether the interaction component is added to the Go or the No-Go action below).

$$W(s_t, a) = \begin{cases} Q(s_t, a) + b + \pi \times V(s_t), & a = \text{Go} \\ Q(s_t, a), & a = \text{No-Go} \end{cases} \quad (3)$$

On each trial, following the reward outcome R_t, the learned action-value of the chosen action was updated according to a standard delta rule:

$$Q(s_t, a) = Q(s_t, a) + \eta(R_t - Q(s_t, a)) \quad (4)$$

Models were fit using a hierarchical framework such that individual participants' parameters were assumed to be drawn from a group-level distribution with a mean and variance that were estimated from the data. Within this overall framework, we compared three models in which group-level parameter distributions were either stable across the two stimulus sets (i.e., same group-level mean and variance parameters with no influence of the affect induction; Model 1), varying between stimulus sets but not across affect-induction conditions (e.g., due to general practice or fatigue effects; Model 2), or varying between stimulus sets in a manner that depended on participants' affect-induction condition (Model 3). That is, Model 1 estimated a single set of parameters per participant (i.e., η, b, π, β), assuming that these parameters were constant across the two stimulus sets. By contrast, for each participant, Models 2 and 3 estimated a set of parameters for the first stimulus set *and* a set of changes to those parameters in the second stimulus set (three[1] additional parameters per participant: $\Delta\eta, \Delta b, \Delta\pi$). Model 3 differs from Model 2 in that it assumed that the group means for these additional parameters differed between the positive and negative affect-induction groups. For Models 2 and 3, this framework allowed us to test whether there was an overall effect of the affect inductions on any parameter by testing whether the mean of the group-level change parameter was credibly different from 0. See Table 1 for a detailed overview of the differing group-level parameters that were fit in each model. Note that this set of parameter changes does not exhaust the set of possible effects of the affect induction that we might have considered in a model; for instance, we might also have considered effects of the affect induction on domain-specific changes in the learning process itself (e.g., Swart et al., 2017). However, we opted to restrict our analyses to affect-related modulations of

[1] Note that the softmax inverse temperature β can be interpreted as an index of the overall goodness-of-fit of a reinforcement learning model. Consequently, to avoid estimating a parameter that might solely reflect differences in model goodness-of-fit we did not allow this parameter to vary between stimulus sets.

parameters within the model specified in Equations 1 to 4, since this is by far the most commonly used model for this task in the literature (Cavanagh et al., 2013; Guitart-Masip et al., 2011, 2012; Moutoussis et al., 2018; Raab & Hartley, 2020).

Model fitting and comparison

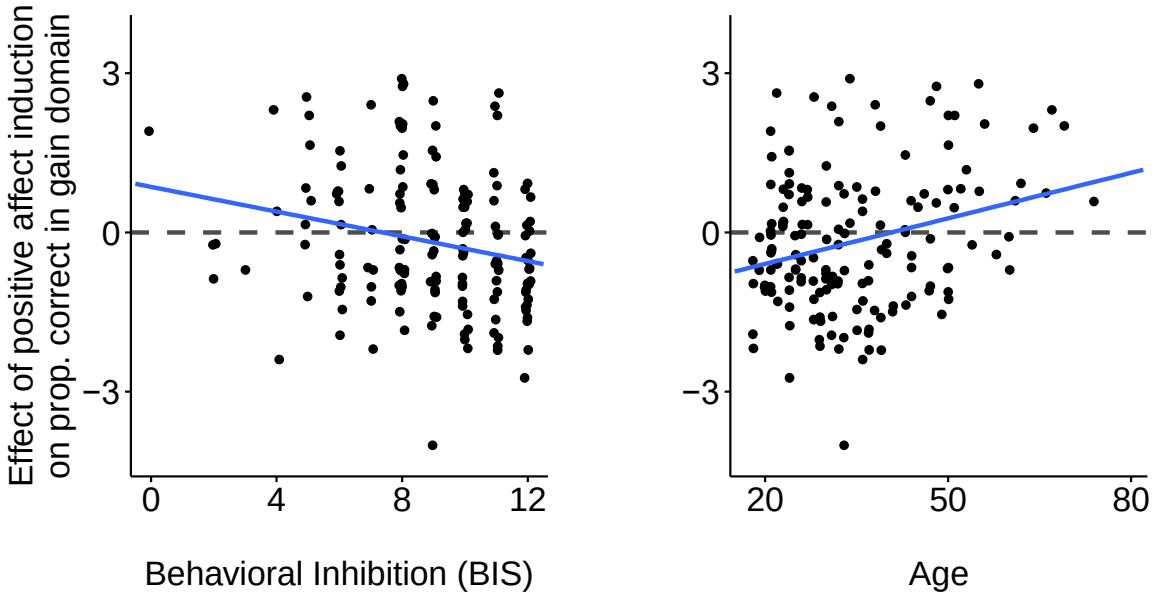
Models were fit in a hierarchical Bayesian framework, using Hamiltonian Monte Carlo as implemented in the software Stan (Carpenter et al., 2017). For each model, four independent chains with independent starting points sampled from the joint posterior distribution. Each chain took 3000 samples from the posterior, with the first 2250 warmup samples discarded to prevent dependence on starting point. The remaining 750 samples per chain were combined to yield 3000 total samples from the joint posterior. There were no divergent transitions in any model, and all chains converged ($\hat{R} < 1.1$ for all parameters). Full Stan code for all models is available in the online project repository.

Models were compared using the Watanabe-Akaike Information Criterion (WAIC), a measure of model fit for hierarchical Bayesian models (Watanabe, 2010). We selected a best-fitting model using pairwise comparisons of the WAIC statistic between each model and the model with the lowest overall WAIC value (i.e., we calculated the pairwise ΔWAIC). Where two models provided a statistically equivalent fit to the data (i.e., where ΔWAIC for a pair of models was less than the standard error of the ΔWAIC for the pair), we selected the simpler of the two models, with model complexity measured as the overall number of free parameters in the model. Parameter inferences were conducted by computing 95% Bayesian Highest Density Intervals (HDIs) across posterior samples (with the exception of exploratory correlational analyses, as discussed below).

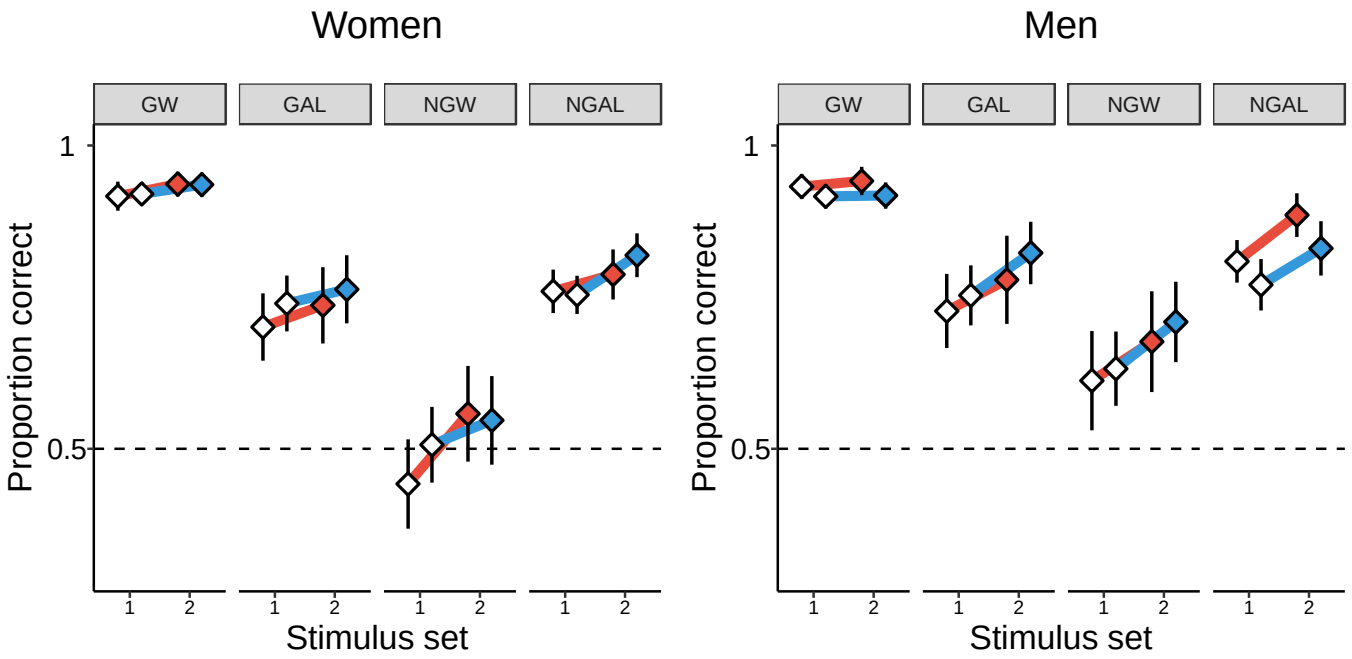
Table S12. Overview of group-level form of parameter distributions used for each of the three models under consideration

Model number	Stimulus set	Parameter			
		Learning rate (η)	Go bias (b)	Pavlovian bias (π)	Softmax inverse temperature (β)
1	1 2	$\eta \sim N(\mu_\eta, \sigma_\eta)$	$b \sim N(\mu_b, \sigma_b)$	$\pi \sim N(\mu_\pi, \sigma_\pi)$	$\beta \sim N(\mu_\beta, \sigma_\beta)$
2	1	$\eta \sim N(\mu_\eta, \sigma_\eta^{set\ 1})$	$b \sim N(\mu_b, \sigma_b^{set\ 1})$	$\pi \sim N(\mu_\pi, \sigma_\pi^{set\ 1})$	
	2	$\eta \sim N(\mu_\eta + \mu_{\Delta\eta}, \sigma_\eta^{set\ 2})$	$b \sim N(\mu_b + \mu_{\Delta b}, \sigma_b^{set\ 2})$	$\pi \sim N(\mu_\pi + \mu_{\Delta\pi}, \sigma_\pi^{set\ 2})$	
3	1	$\eta \sim N(\mu_\eta, \sigma_\eta^{neut})$	$b \sim N(\mu_b, \sigma_b^{neut})$	$\pi \sim N(\mu_\pi, \sigma_\pi^{neut})$	
	2 (pos. affect)	$\eta \sim N(\mu_\eta + \mu_{\Delta\eta}^{pos}, \sigma_\eta^{pos})$	$b \sim N(\mu_b + \mu_{\Delta b}^{pos}, \sigma_b^{pos})$	$\pi \sim N(\mu_\pi + \mu_{\Delta\pi}^{pos}, \sigma_\pi^{pos})$	
	2 (neg. affect)	$\eta \sim N(\mu_\eta + \mu_{\Delta\eta}^{neg}, \sigma_\eta^{neg})$	$b \sim N(\mu_b + \mu_{\Delta b}^{neg}, \sigma_b^{neg})$	$\pi \sim N(\mu_\pi + \mu_{\Delta\pi}^{neg}, \sigma_\pi^{neg})$	

Section S7: Plots of significant demographic moderators, Experiment 2



Supplementary Figure S2. Significant associations between individual differences and affect-induction-related random effects. The vertical axis presents a participant-wise random effect extracted from a mixed-effects regression analysis; this random effect quantifies individual differences in the effect of the positive affect induction on performance for gain-domain stimuli. As such, larger positive values indicate a greater performance improvement for gain-domain stimuli after the positive affect induction (relative to loss-domain stimuli), and negative values indicate a performance deterioration. Points are jittered horizontally to avoid overplotting. Overlaid regression lines indicate the linear association of best fit and its 95% confidence interval. Left panel: there was a significant negative association between self-reported behavioral inhibition (BIS subscale of BIS/BAS) and this random effect (Spearman ρ = -.23, p = .02, FDR-corrected). Right panel: there was a significant positive association between this random effect and participants' age (Spearman ρ = .19, p = .04, FDR-corrected).



Supplementary Figure S3. (D) Proportion correct as a function of stimulus type and stimulus set, averaged separately across women (left) and men (right). Performance improved overall between stimulus set 1 and stimulus set 2. At a group-level, there was no significant difference in the patterns of improvement between participants who received a positive affect induction (red) and those who received a negative affect induction (blue). However, the exact pattern of differences differed significantly between women and men (χ^2 (1) = 7.12, p < .01). Error bars denote the 95% confidence interval of the mean. Note that some very narrow error bars are obscured by condition markers.

Section S8: Model parameter and self-report survey correlation tables

Table S12. Matrix of Spearman correlations between survey self-report measures (total $N = 426$)

	Hypomania (7-up)	Depression (7-down)	BIS	BAS (reward sensitivity)	BAS (drive)	Anxiety (GAD-7)	Worry (PSWQ)	Anhedonia (SHAPS)
Hypomania (7-up)	1	-	-	-	-	-	-	-
Depression (7-down)	.10	1	-	-	-	-	-	-
BIS	- .05	.43	1	-	-	-	-	-
BAS (reward sensitivity)	.25	- .10	.19	1	-	-	-	-
BAS (drive)	.27	- .21	- .07	.52	1	-	-	-
Anxiety (GAD-7)	.21	.69	.42	- .04	- .08	1	-	-
Worry (PSWQ)	.14	.45	.48	.04	- .10	.52	1	-
Anhedonia (SHAPS)	.10	- .32	.05	.42	.35	- .19	- .13	1

Correlations with an absolute value > .14 are statistically significant at $p < .01$ (uncorrected). Correlations with an absolute value > .09 are statistically significant at $p < .05$ (uncorrected).

Table S13. Matrix of Spearman correlations between model parameters (total $N = 426$)

	η	b	π	β	$\Delta\eta$	Δb	$\Delta\pi$
η	1	-	-	-	-	-	-
b	- .12	1	-	-	-	-	-
π	- .46	.30	1	-	-	-	-
β	- .24	- .18	.03	1	-	-	-
$\Delta\eta$	. 10	.01	- .19	- .07	1	-	-
Δb	. 04	.07	.05	- .17	- .06	1	-
$\Delta\pi$	- .15	- .03	.07	.03	- .21	.12	1

Correlations with an absolute value > .14 are statistically significant at $p < .01$ (uncorrected). Correlations with an absolute value > .09 are statistically significant at $p < .05$ (uncorrected).